Predicting Municipal Rental Prices with Advanced Algorithms

Instructor Information:

Resume in brief:

Dr. Pavlos P.

- Scientific Program Director and Lecturer, Institute for Applied Computational Science, Harvard University
- Ph.D., Theoretical Physics, University of Pennsylvania

As Scientific Program Director, he directs the educational program of the institute, overseeing the curriculum of the master's programs in computational science and engineering and data science, teaching core data science courses, and mentoring master's students and advising Ph.D. students in related fields.

His research is at the intersection of astronomy and data science. During his career, he has applied machine learning, deep learning, and statistics to gain knowledge about the cosmos. Currently, he is developing complete and diverse training sets to be used in the application of machine learning to astronomy, using deep transfer learning, generative adversarial networks (GANs), and deep neural networks. His research topics include many areas of machine learning and data science, especially classification, probabilistic graphical modeling, and time series analysis.

1. Program Description

The AI HUB course is an entry-level introduction to data science and it will take you from a basic knowledge of Python to the ability to classify images using deep learning.

Along the way, you will learn data wrangling, cleaning, exploratory data analysis – generating hypotheses and building intuition and the basic concepts of machine learning, including regression, classification, bias, variance, overfitting, and regularization. You will learn how to train deep learning models called Convolutional Neural Networks (CNNs) for image recognition. You will finally learn how to use models pre-trained by others on large image datasets on your own datasets to create models customized to your work (transfer learning).

You will spend a fair bit of time in this course writing your own code and as a result at the end of this course, you will be able to gain a good understanding of machine learning and deep learning, and be able to tackle more advanced tasks and topics in this subject by yourself.

All students must complete a team data science project during the in-class period. The data science project involves the analysis of **apartment rental prediction in Boston**. Students must work in teams (3-5 students) on the project.

The project requires each student team to use python, pandas, sklearn and Keras tool packages to clean up the initial data for. Students are required to combine some algorithms of machine learning to establish and design a simple data science mode. Students need to analyze this built model to make some predictive analysis for the rental market.

2. Prerequisites and Necessary Background

Programming Experience: 1) Experience with Python: functions, classes, modules, NumPy, SciPy; 2) Basic data structures; 3) File I/O.

Statistics Experience: 1) Basics of probability 2) Univariate distributions including the normal and binomial.

3. Working in Groups

Working in project groups may be new for some of you, and can be challenging for all. Spend time at the start of the course learning about prospective group members. Remember that you will have a better group experience if your group is diverse in talents and interest, but united in goals and compatible in work habits. I will do my best to help your group resolve problems, but it typically works best if the group can resolve problems on their own. This will better prepare you to manage group work in "real life."

4. Problem Statement

At some point in time, each of us would have gone through the process of either renting or buying a house.

Investopedia defines residential rental property as follows:

> Residential rental property refers to homes that are purchased by an investor and inhabited by tenants on a lease or rental agreement. Residential real estate can be single-family homes, condominium units, apartments, townhouses, duplexes and so on.

> The term residential rental property distinguishes this class of rental real estate investment from commercial properties where the tenant will generally be a corporate entity rather than a person or family, as well as hotels and motels where a tenant does not live in the property long term."

The renting of residential properties is an economic process that serves a really important role in many communities (especially urban communities) in the United States. First of all, residential leasing is one of the key pathways for property owners and landlords to realize value from their property. In many cases (especially in communities where multi-family homes are common) renting is an essential part of the financial calculus that allows certain classes of people to fulfill their dreams of home/property ownership. A second factor is that residential rentals are an essential mechanism for people to live in a community without the often-onerous burdens of purchasing properties outright. In fact, for a number of the most iconic neighborhoods in the urban United States the essential character of the neighborhood (as well as the properties) would change drastically (in most cases for the worse) if residential renting was abolished.

As such in recent economic climates predicting the short-term leasing values of an apartment or residential property has become a problem of increasing importance. The rise of short-term rental services like AirBnB has increased the urgency of accurate rental value prediction i.e. accurately predicting the (monthly) value of a rental property given various features (including location, condition, size, etc).

5. Project Goals

In this project we'll leverage various sources of data – both aggregate data about municipal features and data about individual rental properties leveraged from commercial and non-commercial sources – to create prediction of property rental revenues for both short-term contexts (e.g. AirBnB) and longer term per month contexts.

The first-step of the project is going to involve scraping individual commercial sites to create a data set of individual property listings. Special care should be taken to create a scalable but respectful scraping process that doesn't put an unnecessary burden on the sties involved. You should also deal with outliers, missing data, and data cleaning. The ultimate output of your scraping process should be a set of tables (think pandas datafames or csvs) representing a dataset of rental properties and their relevant (and uniformly defined/recorded) features

After aggregating your dataset, you may want to generate additional features you think may be relevant (or prune features that aren't of value) using feature engineering and/or some intermediary models. Do rental property pictures add predictive power to any eventual models? [for college students] Does the proximity of a rental property to high paying jobs and/or certain sites of interest help your model(s) in predicting rental revenues? During this stage you'll develop intuition about the data and construct/remove features that may be helpful to your models.

Build (several) models for forecasting both short-term property revenues and monthly long-term revenues. Ideally, you'll use multiple modeling techniques in order to compare your models. Your comparison should include an analysis of the predictive quality as well as an error analysis.

6. Data Resources

We'll provide you with a minimum of data (most of it being in the form of a dataset for short-term property rental info and 'revenue' predictions from AirBnB) but in general you'll be generating your own datasets for this project. The datasets you generate will most likely be composed of data from the following sources:

- Data about individual property listings scraped from commercial sites (think CraigsList, Zillow Trulia, etc). You should be able to get data along the lines of the following:
    - size/square footage
    - rental revenue (in terms of a monthly rent)
    - type of property (e.g. studio, 1/2/3/4 bedrooms, etc)
    - location coordinates (i.e. latitude/longitude)
    - property description
    - amenities (laundry, parking, etc)
    - pictures
- Aggregate data about neighborhoods or municipalities obtained from other (primarily public) sources (average property values, population density, crime rates, school locations, proximity to locations of public interest, transportation characteristics, etc.)
- Data about the property generated (feature engineering) from combinations of the above
- AirBnB dataset containing info collected through AirBnB about short-term property listings and revenues

7. Online Session Schedule

	Weekly Lecture Topic
1	Introduction to the Data Science: Introduction to the course. Brief review of prerequisites. Introduction to statistical learning and data science. Basic concepts. kNN Regression 1.
2	Regression: KNN Regression 2 with examples. Model estimation, error evaluation,

	model fitness and model comparison.
3	Regression: Linear Regression 1. Linear model and model fitting with an example.
4	Classification: Linear Regression 2. Meaning of measurement error, significance of the predictors and bootstrapping.
5	Classification: Multiple Linear Regression, Polynomial Regression, Overfitting, Regularization and Model Selection.