AI Hack 21

Boston Housing Market Challenge

Challenge Description

First studied by Harrison and Rubinfeld (1978), the Boston Housing dataset has been extensively used in testing new machine learning models against existing benchmarks. It is a small dataset with only 506 observations, but is inherently interesting because a lot of things could be studied about this dataset. For this challenge, you are free to pose your own studies and encouraged to use alternative datasets.

You have two options for this challenge. You can either use the housing prices as a response variable and conduct a regression analysis, or focus on the nitrous oxide level and perform a classification task. **You only need to choose one of these two options for your study.** You need not to be comprehensive; the depth of analysis is more important than breaths of the questions posed. You can base your analysis on the available datasets as well as any supplementary datasets you may find.

You may explore one of the sample questions below, or come up with your own variation. Creativity in formulating your own questions is strongly encouraged, although this should not compromise the depth, precision and rigour of your analysis, which will be key performance indicators during assessment.

Data

The original Boston housing dataset can be accessed via the sklearn API

In [ ]:
```
from sklearn import datasets
import pandas as pd
boston_load = datasets.load_boston() boston = pd.DataFrame(boston_load.data,
boston['MEDV'] = boston_load.target
```

A corrected version with town names and spatial information is also available here, which is augmented with longitude and latitude of the observations and corrected for the censoring error. In particular, the censoring error refers to the fact that in the original dataset, the house price is capped at USD 50,000, with values higher than this number set to USD 50,000 (see the description on the page for the corrected dataset: [Ref 1], and also the *Note* section of this page for the original data: [Ref 2]).

References:

[Ref 1] https://nowosad.github.io/spData/reference/boston.html (https://nowosad.github.io/spData/reference/boston.html)

[Ref 2] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html (https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)

Sample Questions:

Regression:

```
[1]. Amongst the given attributes, can you identify any interesting quan
tities that are correlated with the house price, and uncover their stati
stical associations? Can you explain your findings?

[2]. Does the house price show any spatial heterogeneity across the town
s?
You may find the Python library *geopandas* and the *GeoJSON* shapefiles
helpful for visualization.

[3]. A recent study in the Journal of the American Statistical Associati
on (JASA) studied the causal effects of geographical boundaries for hous
e prices in New York City. Can you conduct a similar study?
```

Classification:

```
[1]. Is the nitrous oxide level correlated with any of the explanatory v
ariables? Is this relationship causal?
```

You are reminded that you only need to choose one of regression or classification for your analysis.

Tips and Suggestions:

The following workflow may help:

[1]. Clean the dataset

```
- Impute missing data (if any)
- Do some feature engineering e.g. PCA/t-SNE
```

[2]. Visualise linear correlations.

```
- Can you find any? Compute correlation coefficients
```

[3]. Visualise target distribution

[4]. Regress/classify on all features

```
- Report in sample and out of sample loss
- Report estimated model parameters
- Is spatial adjacency a key information too? How do you add this into y
our model?
```

[5]. Do feature selection/model selection

[6]. Gather your conclusions. What are their implications on economy/policy making?

Contact

Contact : Harrison Zhu (ICDSS, PhD in Modern Statistics and Statistical ML, ICL)
Contact : Xing Liu (ICDSS, PhD in Modern Statistics and Statistical ML, Imperial)
Discord : **https://discord.gg/ymk36q54**

Round 1 Submission

Code Submission

Deadline : Sunday, 21st Feb 2021 at **13:00** UTC/GMT+0
Submission : **https://aihack-2021.devpost.com/**

Report Submission

Deadline : Sunday, 21st Feb 2021 at **13:00** UTC/GMT+0
Submission : **https://aihack-2021.devpost.com/**
Criteria : markdown, pdf, html, or any file formats
that do **not** require special/dedicated tools or software(s)
Tips : Consider these while writing the report.

- What are the goals of your study and why is it important or useful?
- Discuss previous or related work
- Data engineering and processing
- Methodology
- Results: how the results corroborate the assertions in your study.
- Conclusion and discussion, any positive and negative findings.

Presentation Video Submission

Deadline : Sunday, 21st Feb 2021 at **14:00** UTC/GMT+0
Submission : **https://aihack-2021.devpost.com/**
Criteria : Maximum length of 3 minutes

Judging Criteria [Out of 100]

Creativity [15]

Originality of angle of exploration (Interesting questions answered, use of valid alternative dataset(s)

Data Exploration [15]

Quality of techniques used to pre-processed data and to give valuable insights about the dataset(s)

Insight Visualisation [15]

Quality, relevance and effectiveness of visualisations used for exploration and/or analysis

Analytical Techniques [25]

Sophistication and correctness of methods of analysis. Cannot score high if cannot justify method.

Model Validation [5]

Use of metrics in showing performance of analysis.

Interpreting the Result [25]

Ability to interpret the result of the analysis and take a step back to explain the bigger picture. Ability to make a data-driven "business" decision.