Lab. – Regression :

House Sale Price Prediction Challenge

Yuan-Fu Liao

National Taipei University of Technology

房價預測 (回歸模型)

- 任務
  - 用train.csv跟valid.csv訓練模型（一行是一筆房屋交易資料的紀錄，包括id, price與21種房屋參數）
  - 將test.csv中的每一筆房屋參數，輸入訓練好的模型，預測其房價
  - 將預測結果上傳到Kaggle（從“Submit Predictions”連結）
  - 看系統幫你算出來的Mean Abslute Error（MAE，就是跟實際房價差多少，取絕對值）分數夠不夠好？
  - 嘗試改進預測模型

- 討論
  - 資料分析
  - 做法
  - 程式寫法
  - 結果分析
  - 檢討與改進

作業進行方式

- 使用Github與Kaggle的教室平台
  - 把結果放到Kaggle（submission）上排名
  - 把程式與報告放到Github（至少要有readme.md（報告）, train.sh，test.sh，requirement.txt與報告）

- 以邀請碼（另外通知）連結申請GitHub帳號，並以“學號+姓名”，當你的帳號名稱（名稱需與Kaggle的帳號相同
  - 建帳號時要跟課程學生名單上你的學號連結

- 以邀請碼（另外通知）連結申請Kaggle帳號，並以“學號+姓名”，當你的帳號名稱帳號（名稱需與Github的帳號相同）：
  - 與你的github帳號關連

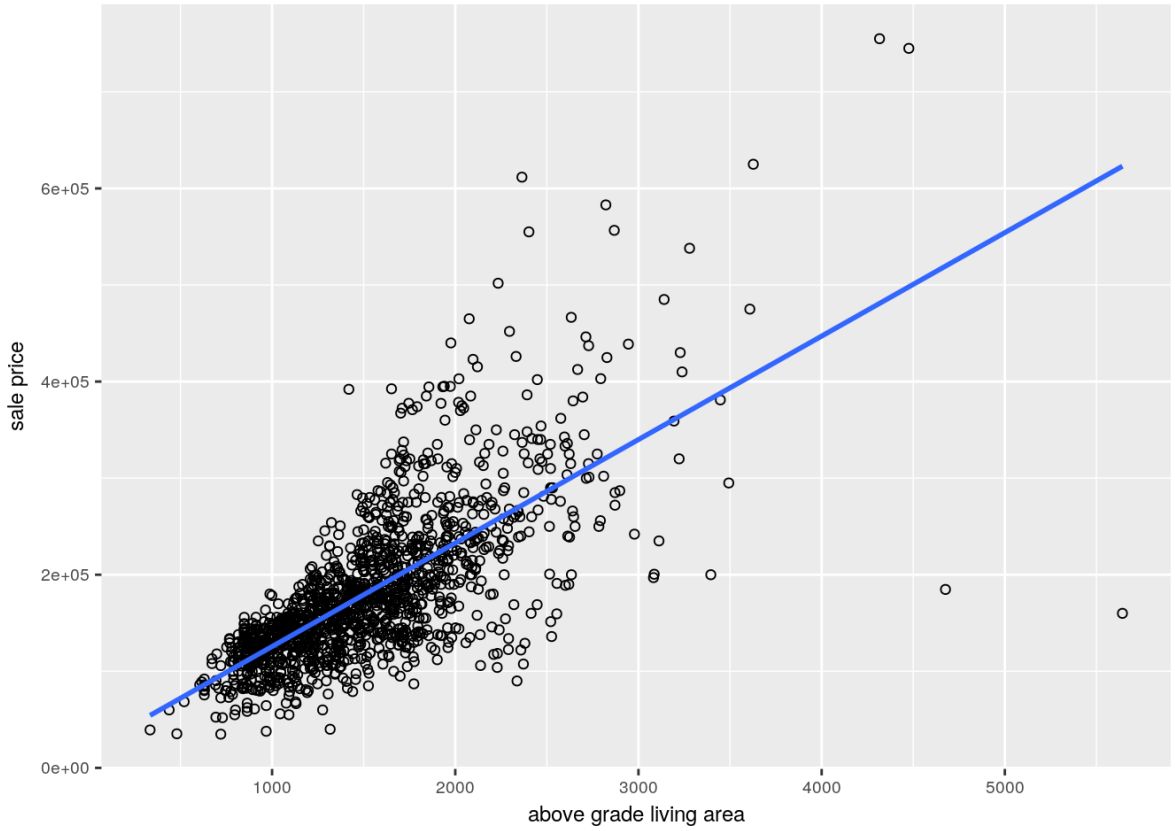
房屋交易資料

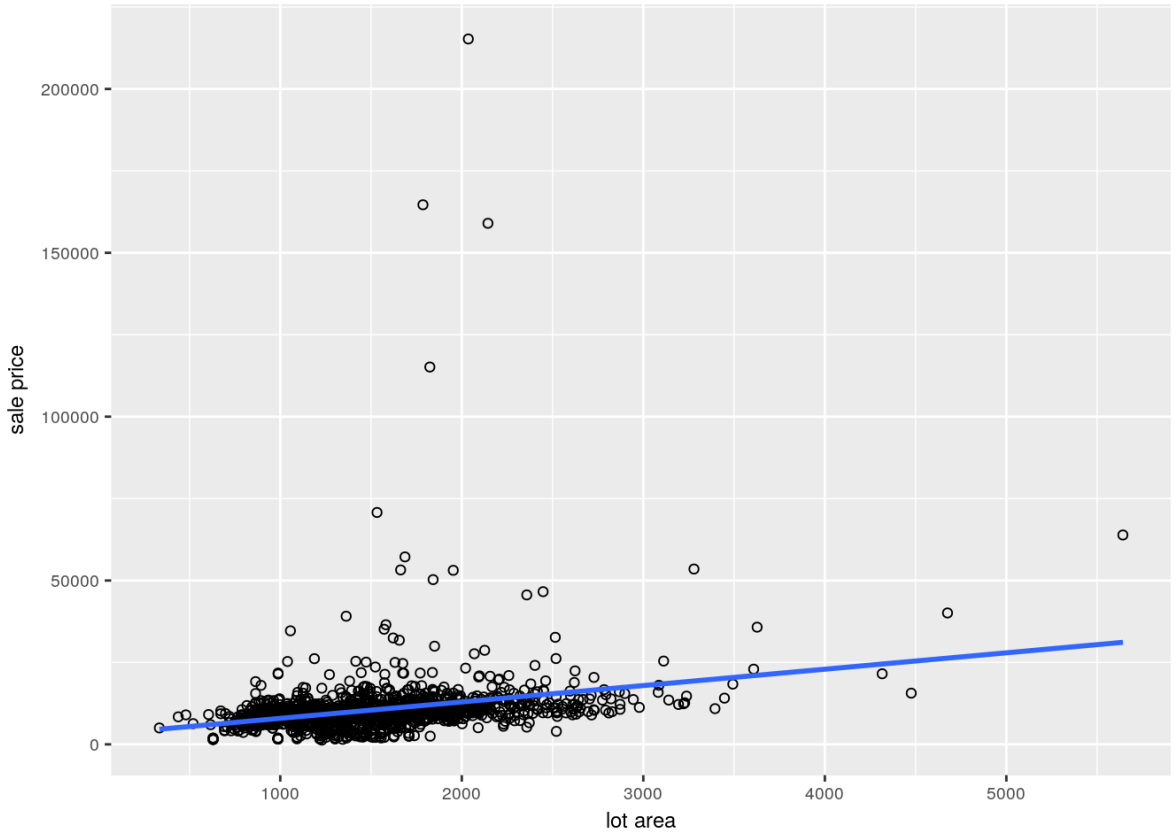
	price	bedrooms	bathrooms	sqft_living	sqft_lot	waterfront	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
0	221900.0	3	1.00	1180	5650	0	1180	0	1955	0	98178
1	538000.0	3	2.25	2570	7242	0	2170	400	1951	1991	98125
2	180000.0	2	1.00	770	10000	0	770	0	1933	0	98028
3	604000.0	4	3.00	1960	5000	0	1050	910	1965	0	98136
4	510000.0	3	2.00	1680	8080	0	1680	0	1987	0	98074

房屋屬性

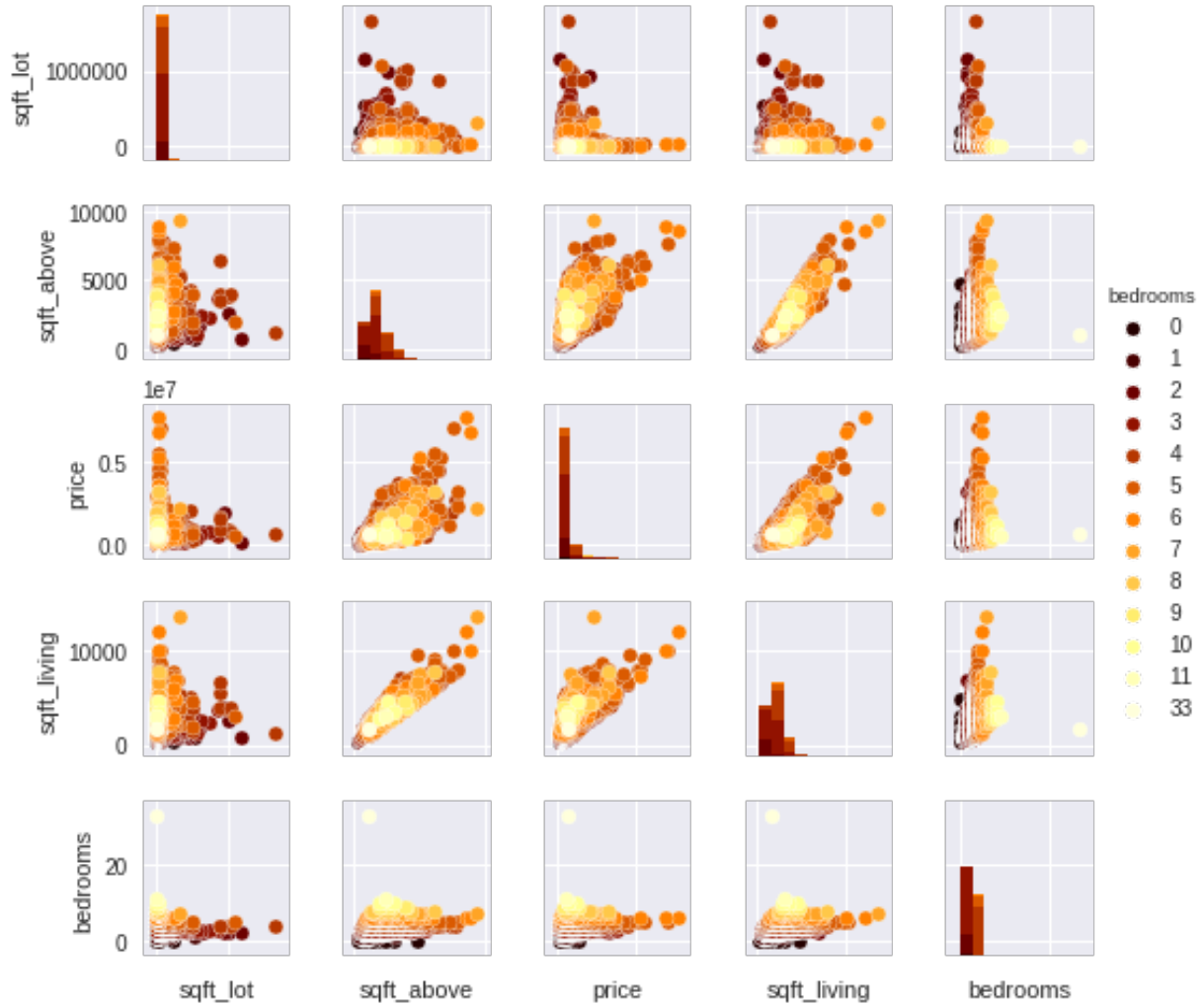
id	a notation for a house	Numeric
year	date house was sold	String
month	date house was sold	String
day	date house was sold	String
price	Price is prediction target	Numeric
bedrooms	Number of Bedrooms/House	Numeric
bathrooms	Number of bathrooms/bedrooms	Numeric
sqft_living	square footage of the home	Numeric
sqft_lot	square footage of the lot	Numeric
floors	Total floors (levels) in house	Numeric
waterfront	House which has a view to a waterfront	Numeric
view	Has been viewed	Numeric
condition	How good the condition is (Overall)	Numeric
grade	overall grade given to the housing unit	Numeric
sqft_above	square footage of house apart from basement	Numeric
sqft_basement	square footage of the basement	Numeric
yr_built	Built Year	Numeric
yr_renovated	Year when house was renovated	Numeric
zipcode	zip	Numeric
lat	Latitude coordinate	Numeric
long	Longitude coordinate	Numeric
sqft_living15	Living room area	Numeric
sqft_lot15	lotSize area	Numeric

Correlation between a Feature and the Price

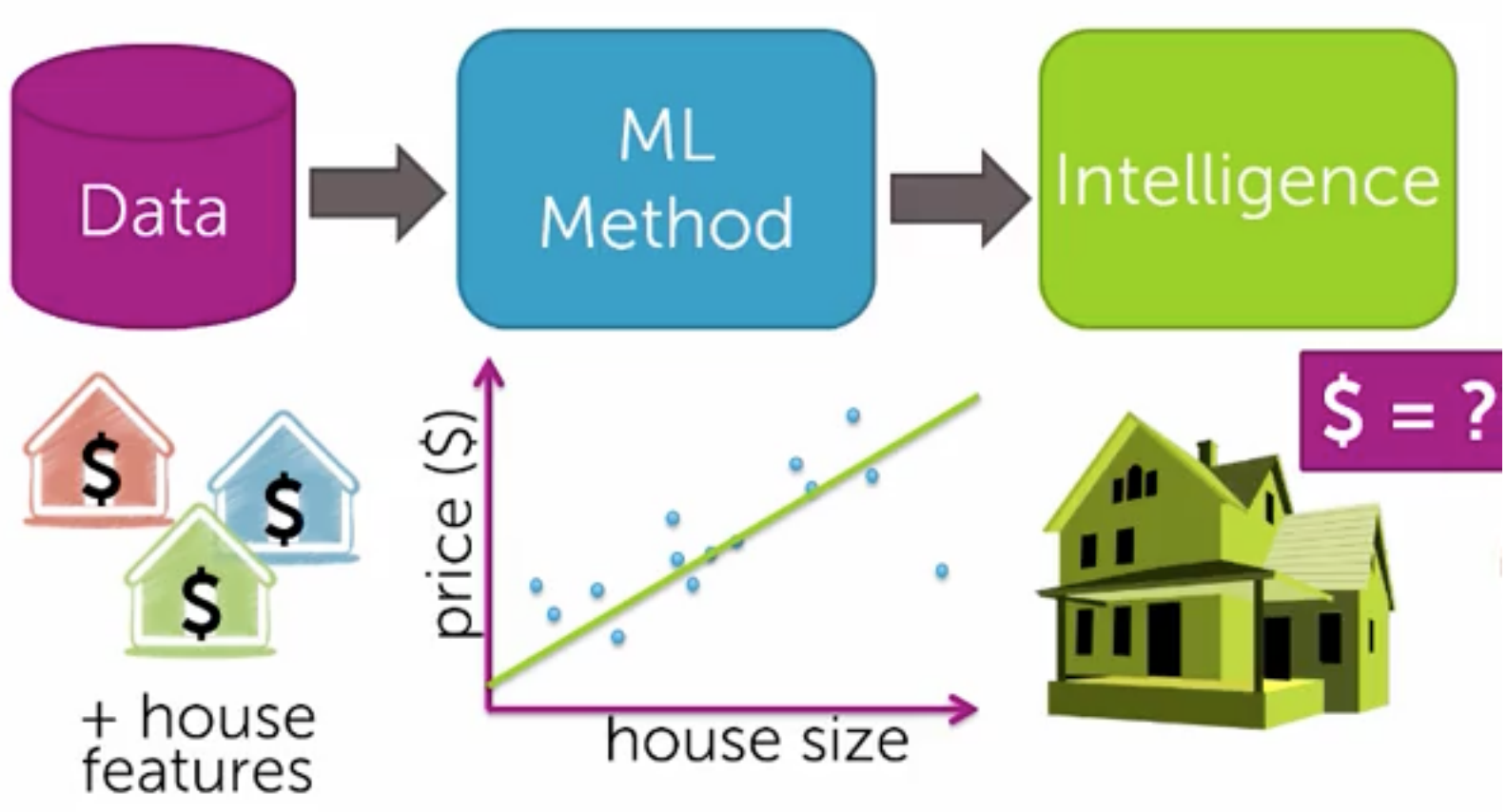




Correlation between Features

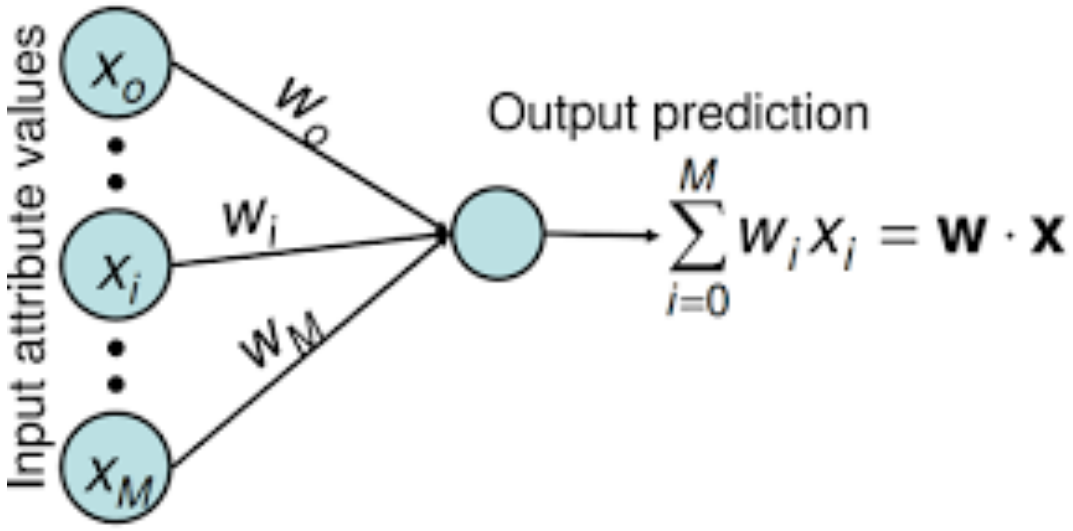


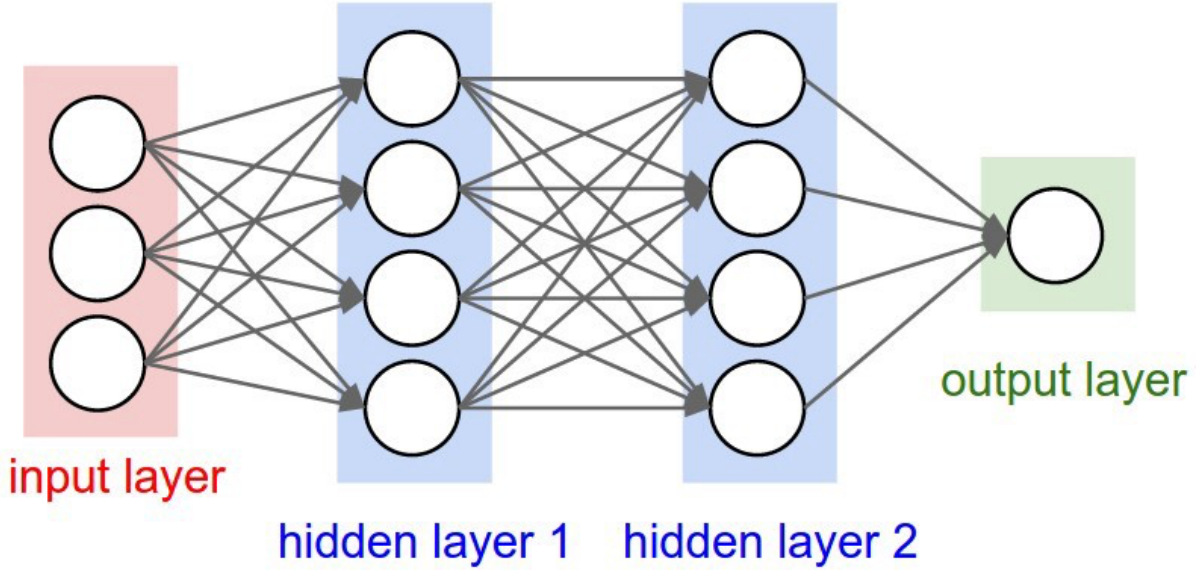
以機器學習演算法訓練回歸模型



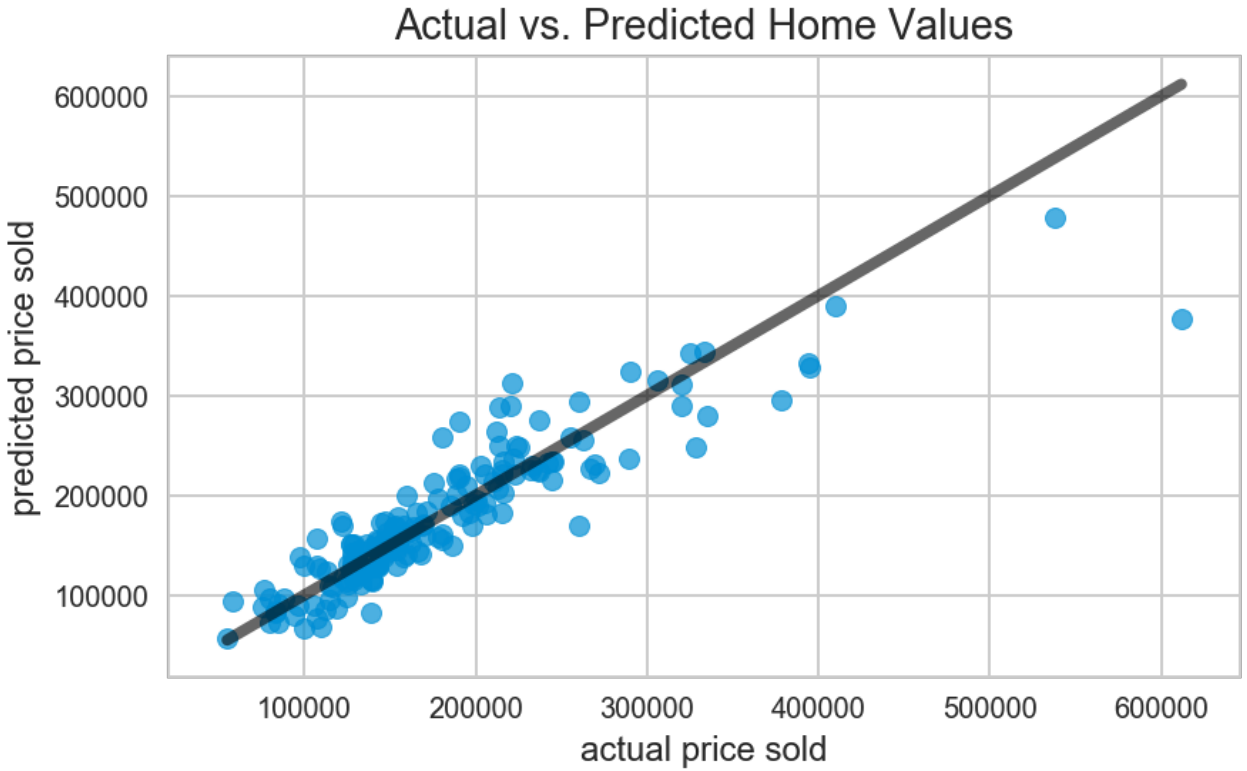
Neural Network-based Regression

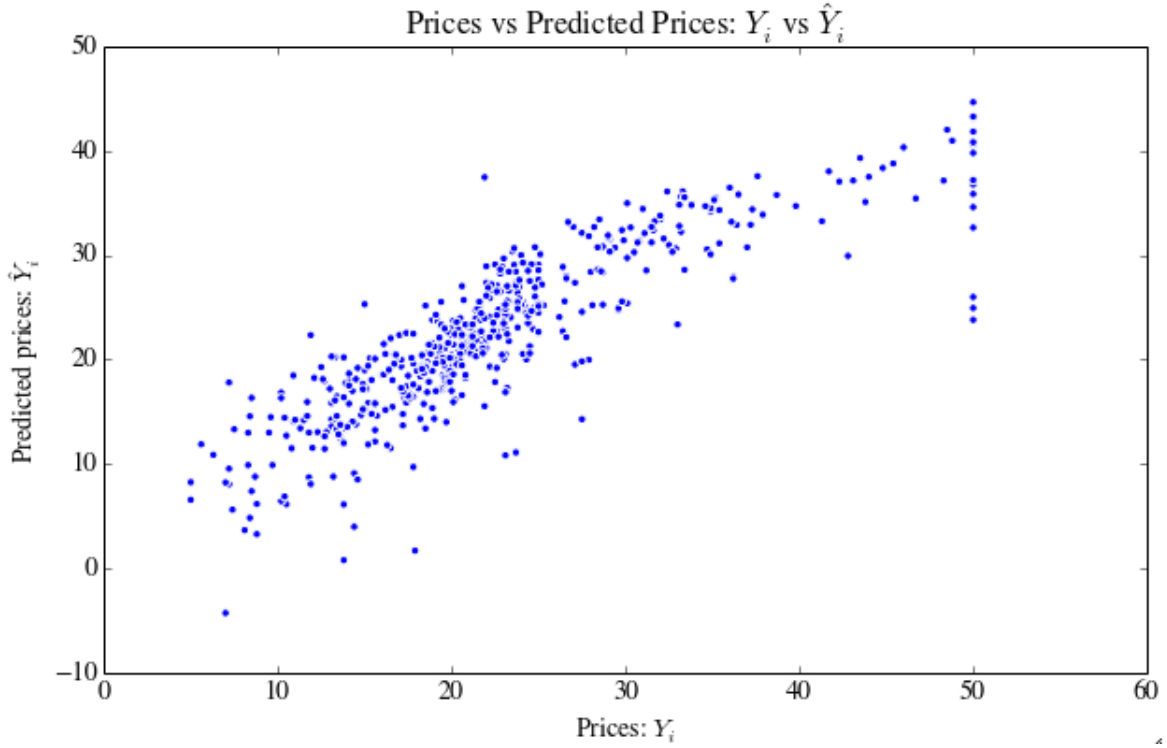
- Linear vs. nonlinear model





Result Analysis



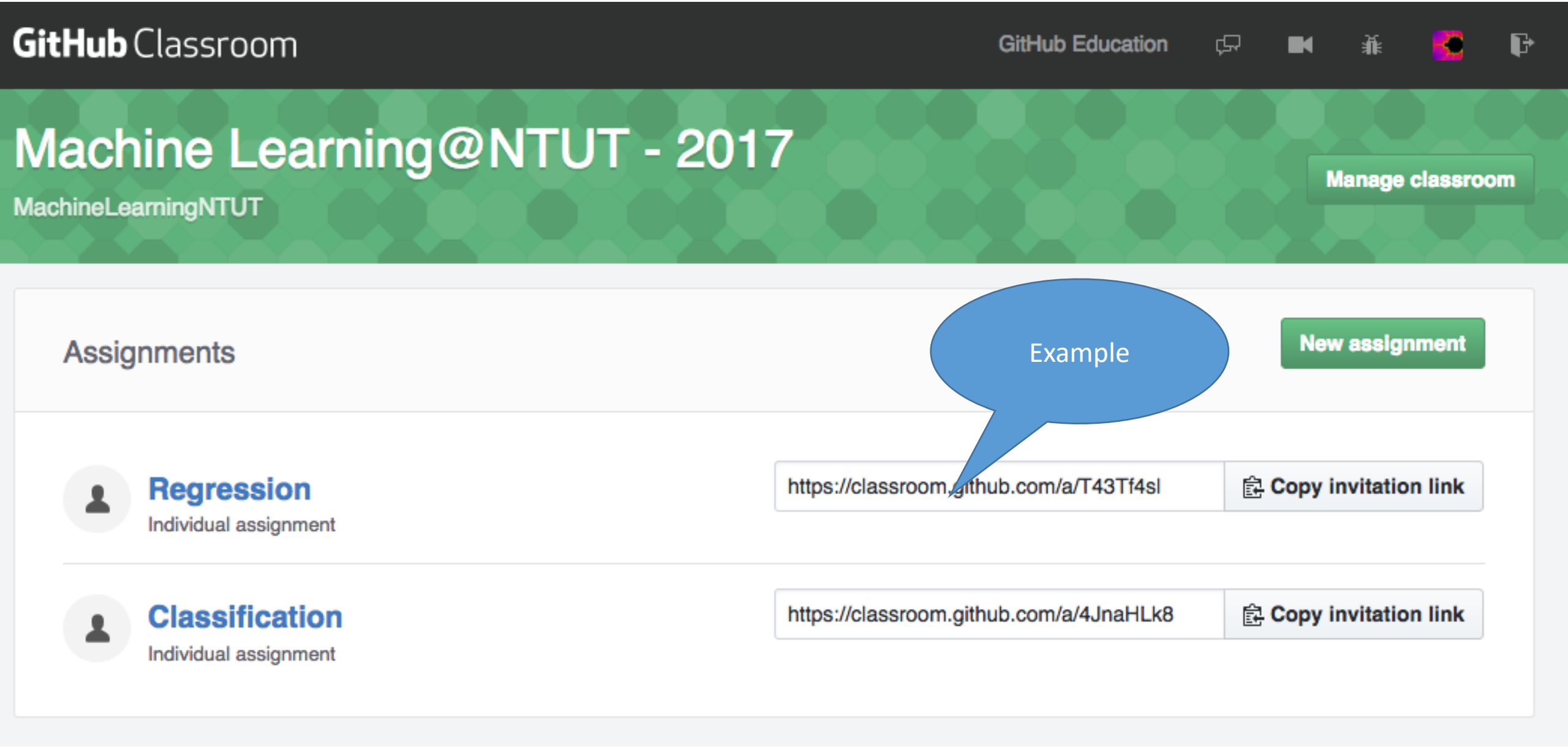


Getting Started With

GitHub

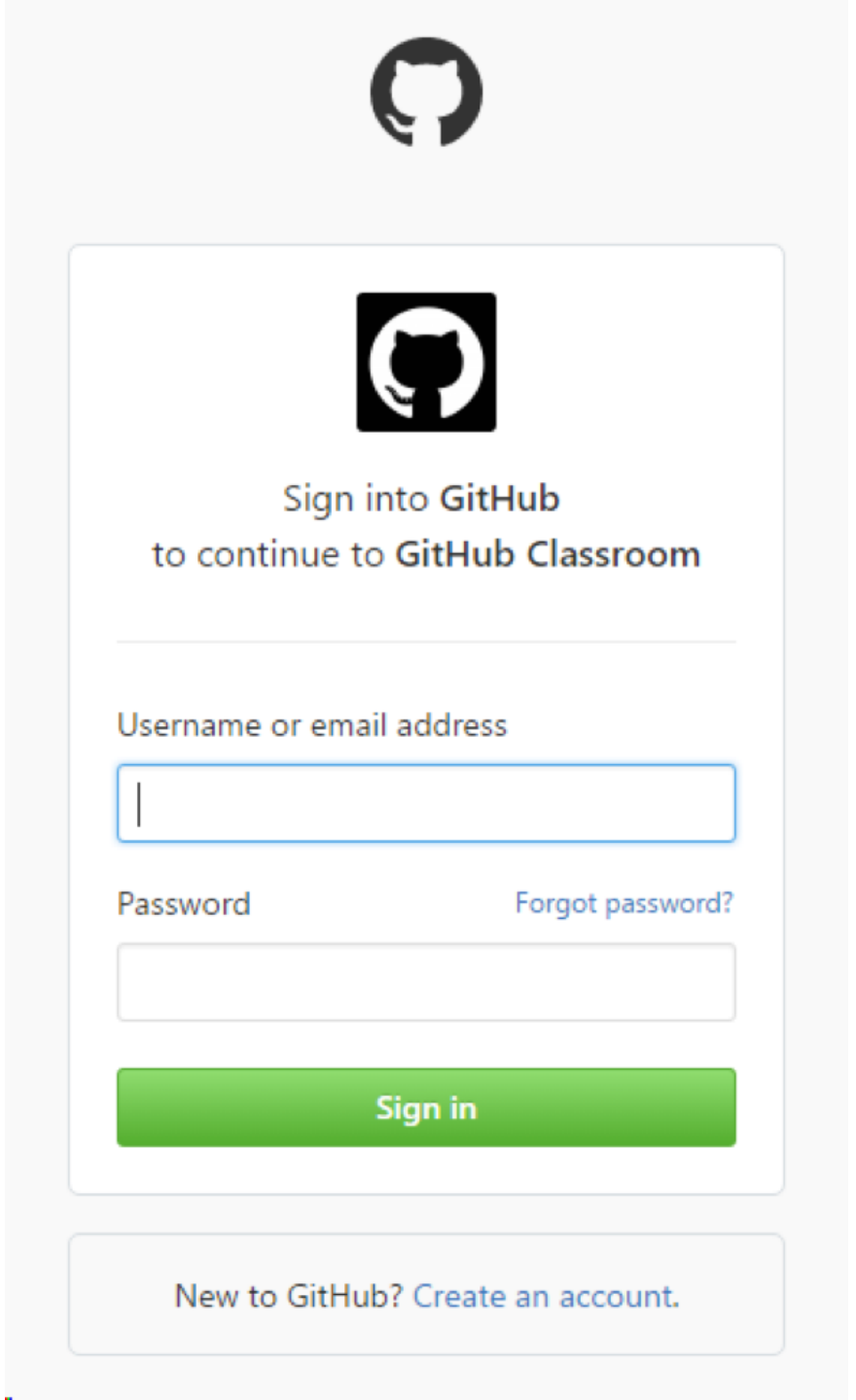
LAU

GitHub Classroom

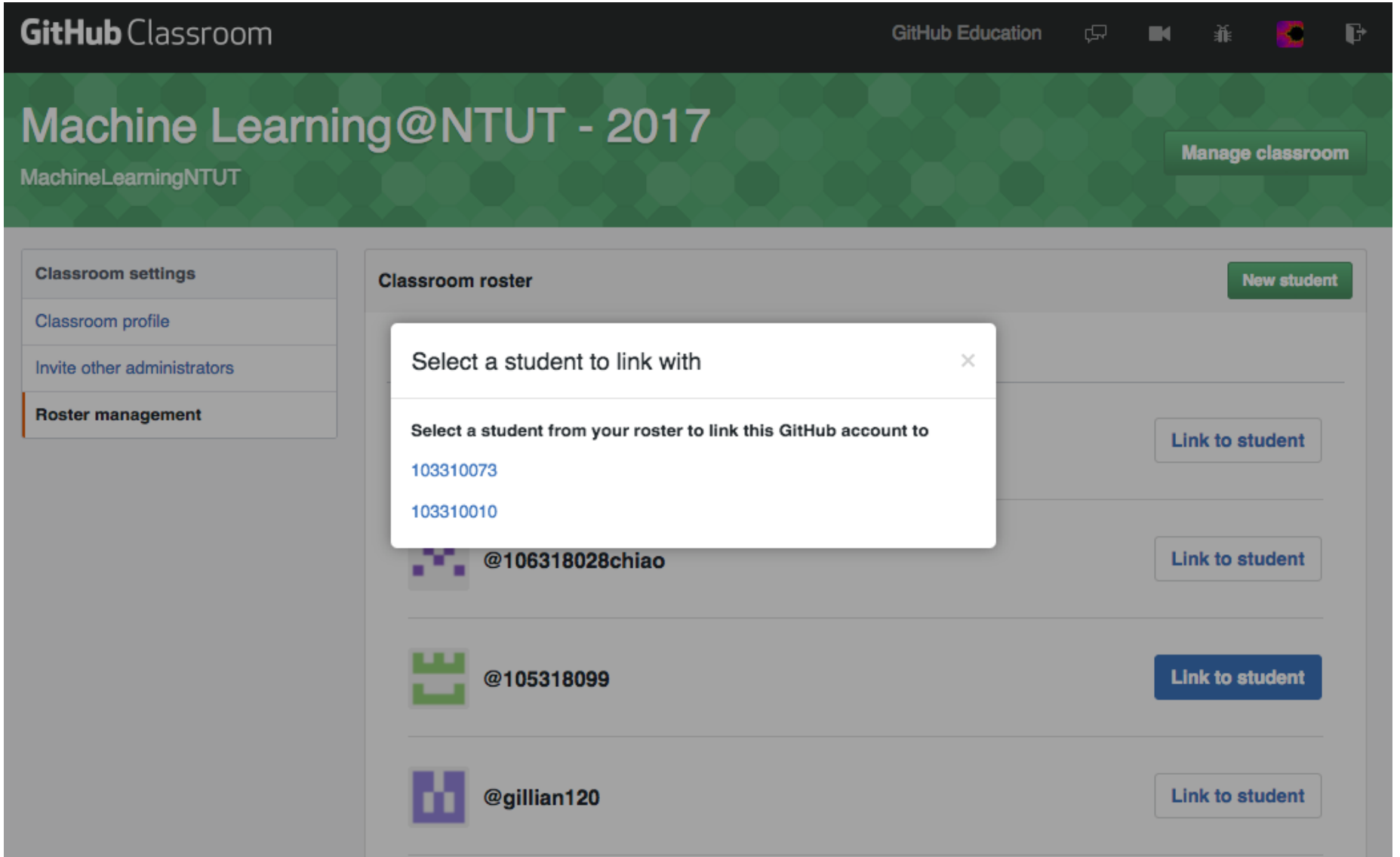

GitHub Classroom
GitHub Education
Machine Learning@NTUT - 2017
MachineLearningNTUT
Manage classroom
Assignments
Example
New assignment
Regression
Individual assignment
https://classroom.github.com/a/T43Tf4sl
Copy invitation link
Classification
Individual assignment
https://classroom.github.com/a/4JnaHLk8
Copy invitation link

Create GitHub Classroom Account

- 以“學號+姓名”當你的帳號名稱



連結課程學生名單（學號）


GitHub Classroom
GitHub Education
Machine Learning@NTUT - 2017
MachineLearningNTUT
Manage classroom
Classroom settings
Classroom profile
Invite other administrators
Roster management
Classroom roster
New student
Select a student to link with
Select a student from your roster to link this GitHub account to
103310073
103310010
Link to student
@106318028chiao
Link to student
@105318099
Link to student
@gillian120
Link to student

MachineLearningNTUT / Regression-yfliao Private

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Settings Insights

Quick setup — if you've done this kind of thing before

Set up in Desktop or HTTPS SSH https://github.com/MachineLearningNTUT/Regression-yfliao.git

We recommend every repository include a README, LICENSE, and .gitignore.

...or create a new repository on the command line

```
echo "# Regression-yfliao" >> README.md
git init
git add README.md
git commit -m "first commit"
git remote add origin https://github.com/MachineLearningNTUT/Regression-yfliao.git
git push -u origin master
```

...or push an existing repository from the command line

```
git remote add origin https://github.com/MachineLearningNTUT/Regression-yfliao.git
git push -u origin master
```

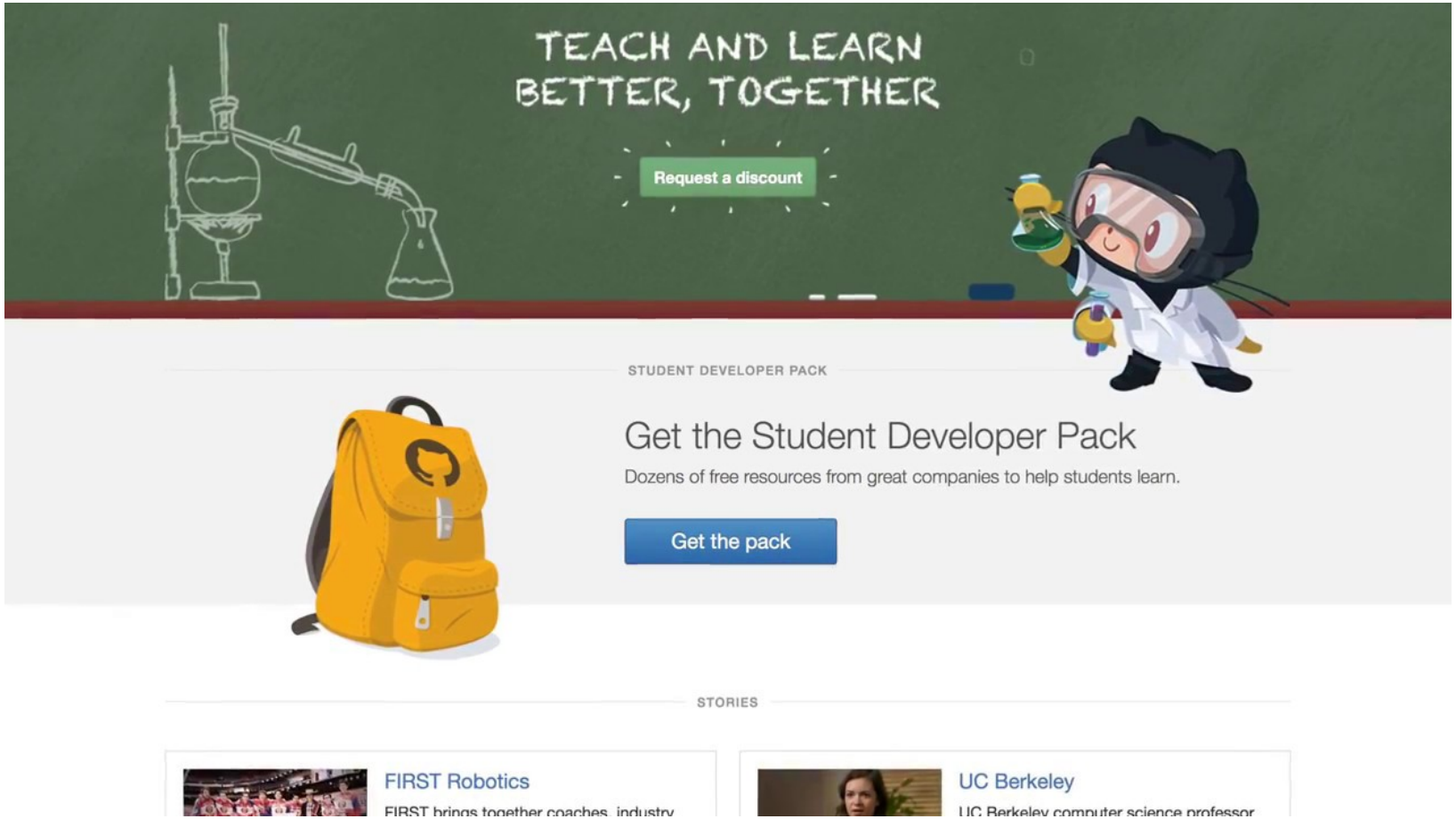
...or import code from another repository

You can initialize this repository with code from a Subversion, Mercurial, or TFS project.

Import code

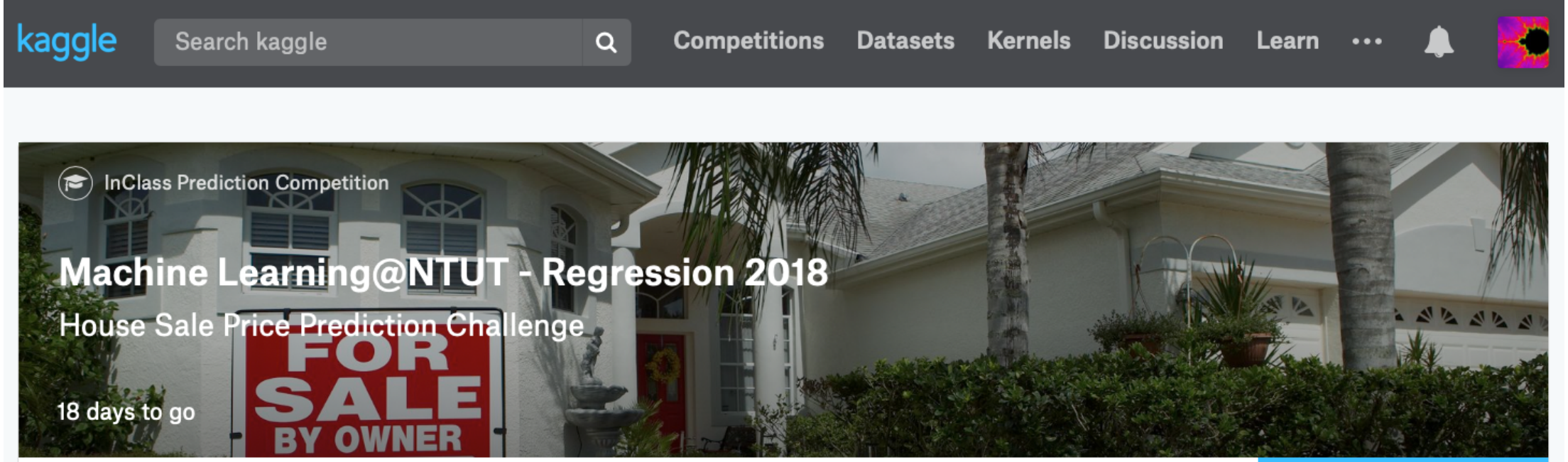
ProTip! Use the URL for this page when adding GitHub as a remote.

Student Developer Pack

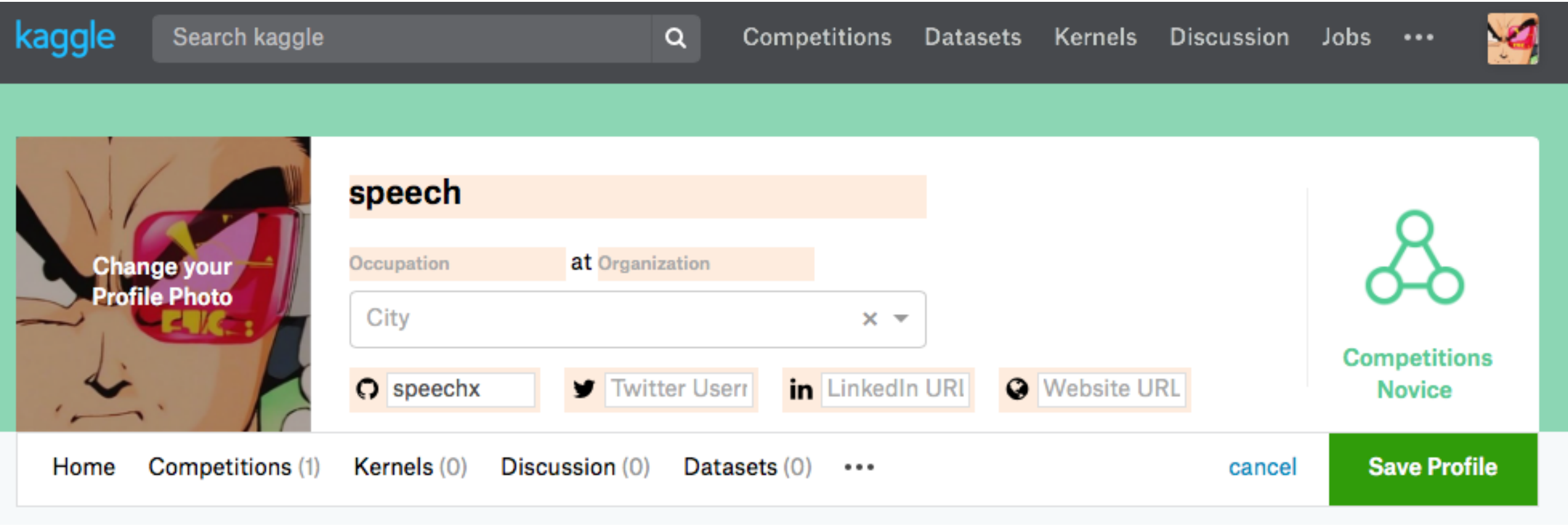


Getting started with Kaggle

House Prices Competition


kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Learn
InClass Prediction Competition
Machine Learning@NTUT - Regression 2018
House Sale Price Prediction Challenge
18 days to go
FOR SALE BY OWNER
Overview
Data
Kernels
Discussion
Leaderboard
Rules
Team
Host
My Submissions
Submit Predictions
Host Controls
Settings
Images
Privacy
Evaluation
Sandbox Submissions
All Submissions
Example
Privacy
limited
Participation of this competition is restricted to those
access to the link below. Note that anyone is still able to view the details of this co
etition.
URL for Sharing
https://www.kaggle.com/t/d071d893
b3d46328e94caa0d1
This is a URL that can be shared around and re-used. Anyone who visits this link will be able to participate in the competition.

跟GitHub帳號連結



Overview Edit

Description

Evaluation

+ Add Page

Please use the given house features to predict its sale price.

The given database for training a sale price predictor contain house sale prices for somewhere out there with 19 house features plus the price and the id columns.

此次作業是使用回歸模型做房價預測，你們要做的是：

1. 用train.csv跟valid.csv訓練模型（一行是一筆房屋交易資料的紀錄，包括id, price與19種房屋參數）
2. 將test.csv中的每一筆房屋參數，輸入訓練好的模型，預測其房價
3. 將預測結果上傳（從"Submit Predictions"連結）
4. 看系統幫你算出來的Mean Abslute Error（MAE，就是跟實際房價差多少，取絕對值）分數夠不夠好？
5. 嘗試改進預測模型

程式要放到Github Classroom

報告要包括：

1. 做法說明
2. 程式方塊圖與寫法
3. 畫圖做結果分析
4. 討論預測值誤差很大的，是怎麼回事？
5. 如何改進？

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | Host | My Submissions | Submit Predictions

Data Description

Edit

This database contains house sale prices for somewhere out there with 21 house attributes plus the price and the id columns.

File descriptions.
1. train-v3.csv - the training set.
2. valid-v3.csv - the validation set.
3. test-v3.csv - the test set sample.
4. Submission.csv - a sample submission file in the correct format.
5. metaData.csv - supplemental information about the data.

Data (725 KB)

API | `kaggle competitions download -c ntut-ml-regressi…` | ? | Download All

Data Sources

+ New Version

- metadata.csv 22 x 3
- sampleSubmission.c... 6485 x 2
- test-v3.csv 6485 x 22
- train-v3.csv 13.0k x 23
- valid-v3.csv 2161 x 23

About this file

Edit

Help us describe this file

Columns

Edit

- A id
- A a notation for a house
- A Numeric

Overview Data Kernels Discussion Leaderboard Rules Team Host My Submissions **Submit Predictions**

Make a submission for **yfliao**

You have 10 submissions remaining today. This resets 18 hours from now (00: 00 UTC).

Step 1
Upload submission file

Upload Submission File

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 6485 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

Styling with Markdown supported

Briefly describe your submission.

Make Submission

Example Codes

- Tensorflow
  - https://www.tensorflow.org/tutorials/keras/basic_regression
- Keras
  - https://github.com/sunyam/HousePrices_Regression
- Kaggle
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- References
  - **Regression Tutorial with the Keras Deep Learning Library in Python**
    - http://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/
  - **Model Evaluation and Validation: Predicting Boston Housing Prices**
    - https://olegleyz.github.io/boston_housing.html

Example: Boston House Price Dataset

ta Set Characteristics:	Multivariate	**Number of Instances:**	506	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	14	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	No	**Number of Web Hits:**	328614

Attribute Information:
1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

- Download: Boston house price dataset