大数据机器学习模型在经济学的应用

Machine Learning for Economists （Part II）

葛雷

中国人民大学经济院

2022 年 5 月 28 日

中國人民大學
RENMIN UNIVERSITY OF CHINA

Jupyter Lab

ML Project A: Income vs GPA

ML Project B: Car Auction (Real Data)

Through two simple projects, you can understand:

- Why machine learning is good at economic & financial predictions (也就是Machine Learning为什么好？)
- Why machine learning is not a blackbox & easy to interpret (Machine Learning为什么方便应用、而且结果一目了然)
- The end-to-end process of the quant modeling in the financial corp. (公司内真实的建模)

Jupyter

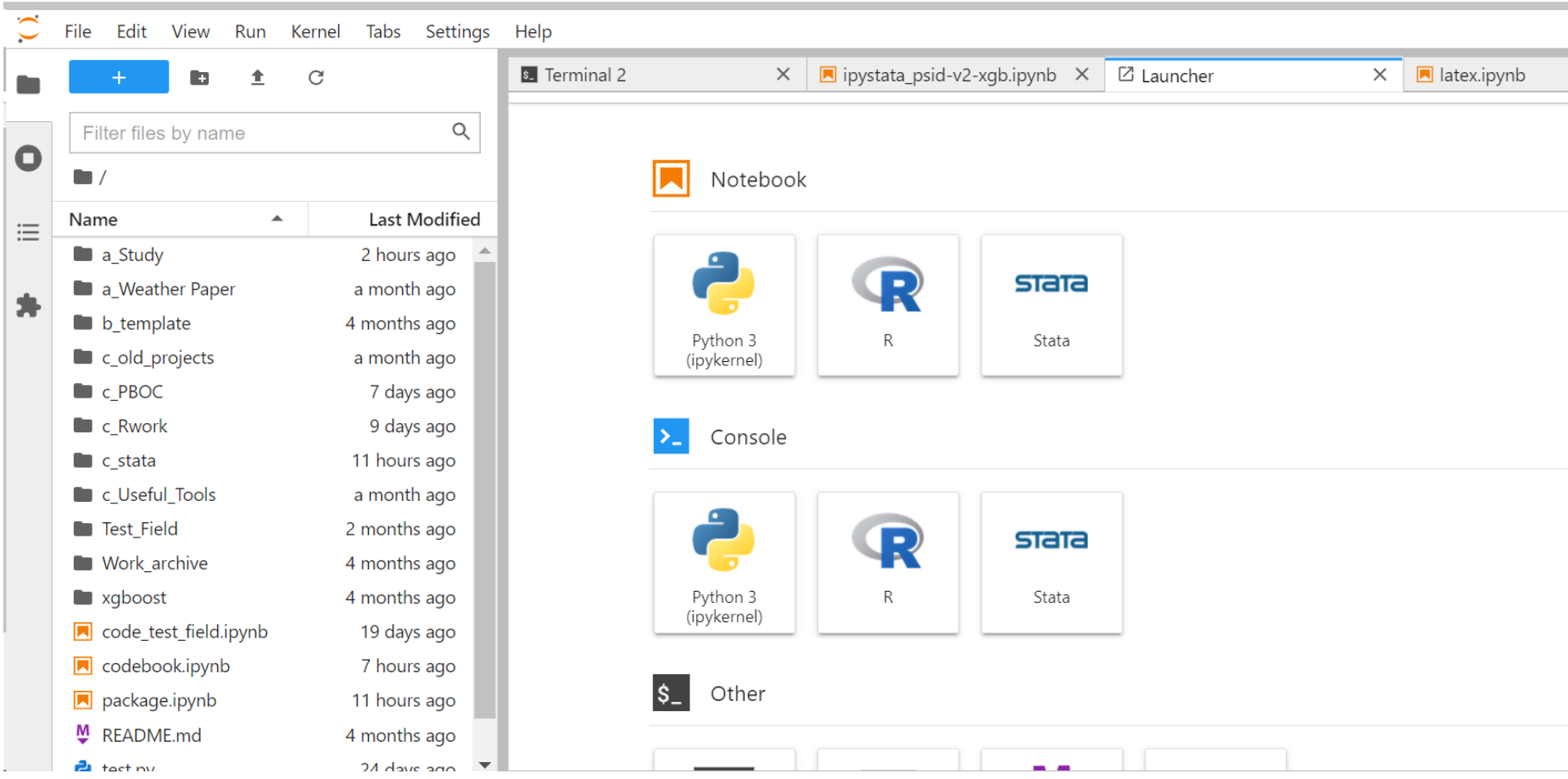


图: Jupyter操作平台

Python packages

Install xgboost、scikit-learn、pandas、Tensorflow **Please turn to the StackOverflow & search engines for all the bugs and problems for the tech issues**

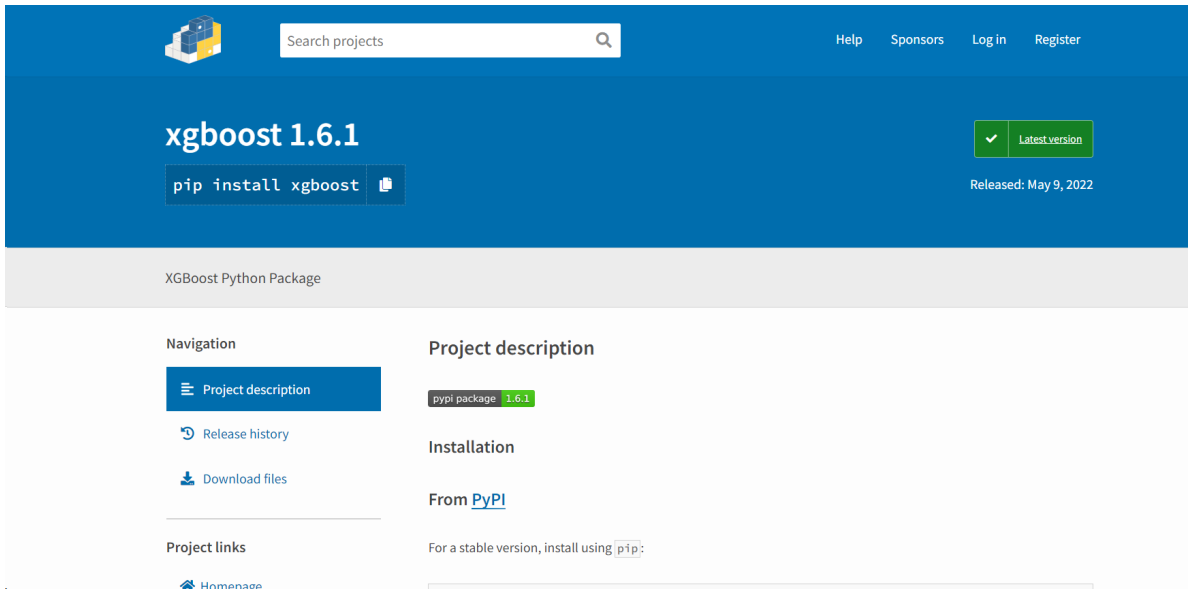
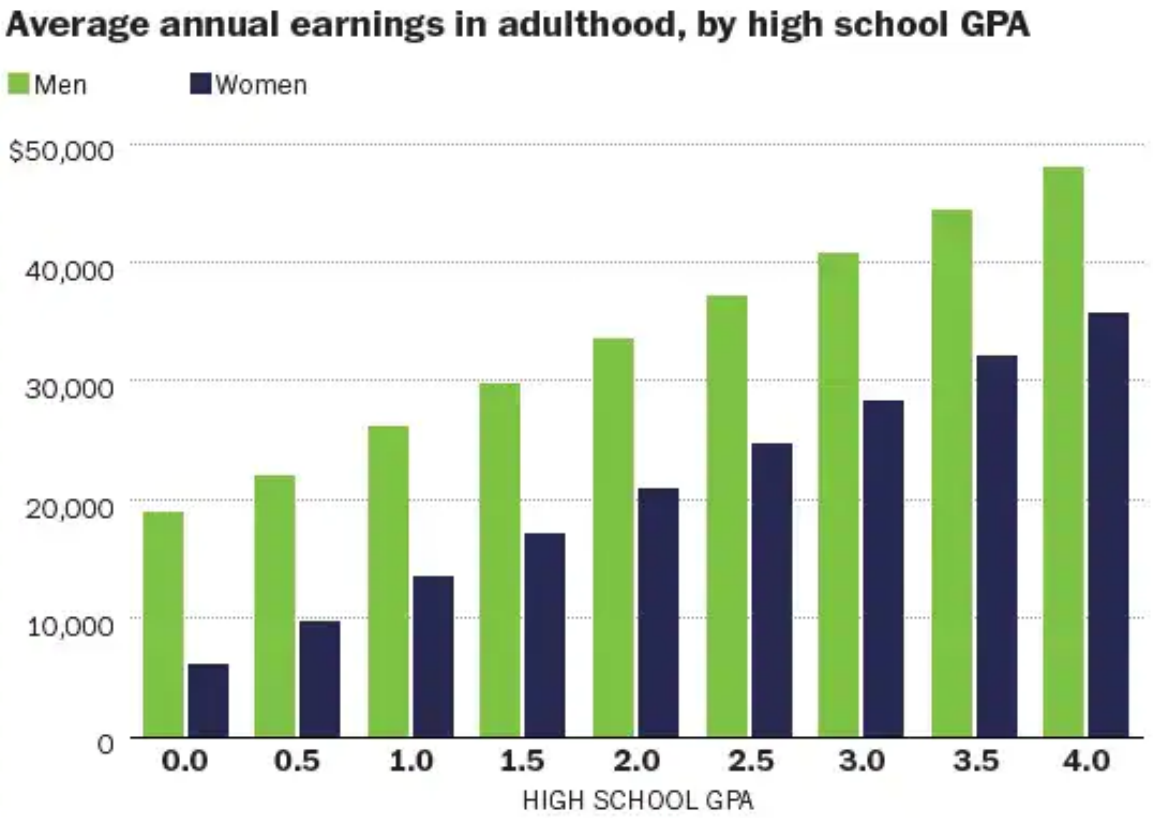


图:

ML项目 A： Income vs GPA

- we use a simple economic model to study the various of trending machine learning models
- through this simple example, you can clearly see why machine learning models are good at predictions

Build a simple economic model



Simple framework between income & GPA

Assume:

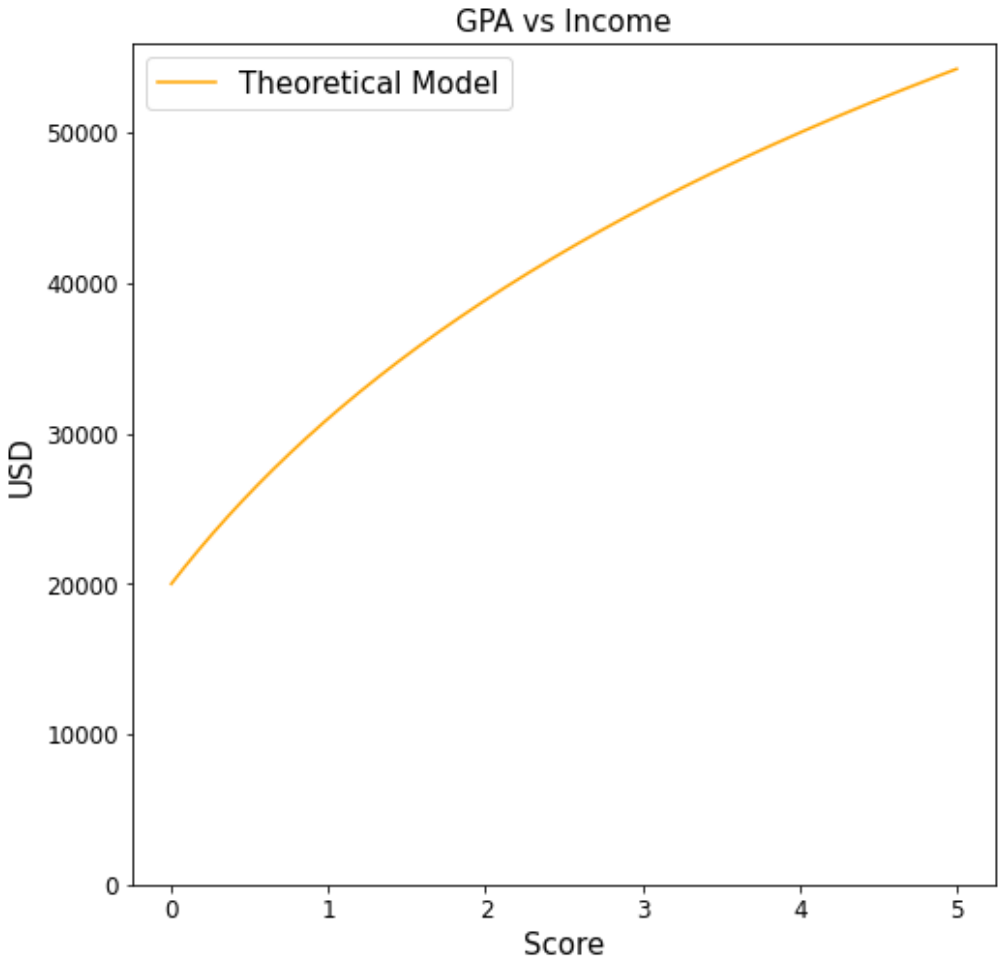
$$income_i = \alpha \log(gpa_i + constant) + \mu_i,$$

where i is the person ID, $income_i$ is the income of the person i and gpa_i is her GPA during the college study. Assuming, μ_i follow a normal distribution $N(0, \sigma)$. (Noted:α and *contant* are predetermined values)

Simple framework between income & GPA

Assume:

$$income_i = \alpha \log(gpa_i + constant) + \mu_i,$$



Simulate Mocking Dataset

Then, We use computer to simulate 10,001 persons from this framework above:

Choices of quant models

1. 简单统计（Simple Statistics）：The Han Dynasty and the Roman Empire were some of the first states to extensively gather data on the size of the empire's population, geographical area and wealth.

2. 线性模型(Linear Model)： Legendre (1805) and Gauss (1809) used the linear regression for the prediction of planetary movement.

3. 深度学习(Deep Learning)： Warren McCulloch and Walter Pitts (1943) opened the subject by creating a computational model for neural networks.

Models Comparison（模型的比较与演化）

Please turn to the **GPA_Income.ipynb**:

1. simple average
2. average by groups (actually it is the decision tree)
3. linear regression
4. XGBoost
5. Deep Learning

Why we need to tune the hyperparameters?

- hyperparameters are important part of the machine learning models
- hyperparameters decide the structure of the models
- the choice of the hyperparameters has huge impact on the model's performance

Example: History vs the learning rate

Let turn to the history example from the last class:

- 人类从历史学到的唯一的教训，就是人类没有从历史中吸取任何教训。(We learn from history that we do not learn from history.)
- 用机器学习算法的角度总结黑格尔的话，就是人类社会的学习效率太低太慢。也就是Learning rate （η） is too low。
- Learning Rate是机器学习模型中的一个重要的超参数(Hyperparameters)，控制着机器学习算法的学习速度
- So, we just should choose a extremely large value of learning rate? Nope.

ML Project B: Car Auction (Real Data)

We already learned the how machine learning works and why we need to use machine learning model for the economic predictions from the simulated project A. Here we turn to a real data example to go through the end to end modeling process.

Kaggle datasets & Kaggle API

Please turn to:

- the "1.downdata_kaggle.ipynb"

Data Analysis

Please turn to:

- the "2.data_analysis.ipynb"
- the "2.data_analysis-auto.ipynb"

Quant Modeling

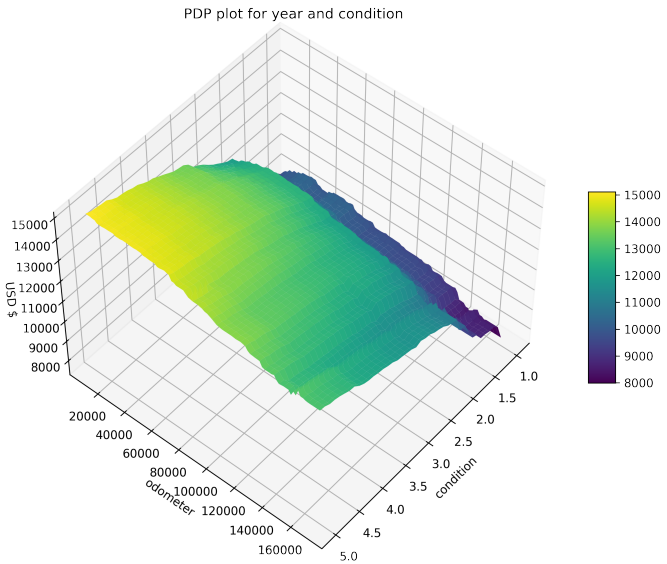
Please turn to:

- the "3.ML modeling.ipynb"

Interpretable Machine Learning

Please turn to:

- the "4.Interpretable ML.ipynb"



Reference

- https://en.wikipedia.org/wiki/History_of_statistics