AI & ML for Data Scientists

Class 5: Build a Model

Lei Ge (葛雷)

Quant RUC (人大量化)

October 15, 2025

Why we need model?

Why the companies will recruit you to build model?

- Prediction and Decision-Making for the Business
- Efficiency
- Precision
- $\rightarrow$ Money for You and Your Company

How to build a professional Model for business?

- For the data science, we usually have 6 steps to build a model:
  - Problem Definition and Data Collection
  - Data Preparation □
  - Model Development
  - Model Evaluation and Optimization □
  - Model Deployment
  - Continuous Improvement

Problem Definition and Data Collection

- **Understand the problem**: Define the objective and scope of the machine learning task.
- **Collect data**: Gather relevant data from various sources.

Data Preparation

- **Preprocess data**: Clean and prepare the data for analysis.
- **Split the data**: Divide into training, validation, and test sets.
- **Normalization/Scaling**: Ensure data is on a consistent scale.
- **Feature engineering**: Create meaningful features from raw data.

Preprocess data

- Drop bad data
- Winsorize data
- Missing value imputation

Model Development

- **Select an algorithm**: Choose a suitable machine learning algorithm.
- **Develop baseline models**: Create simple models for comparison.
- **Train the model**: Use the training data to fit the model.

Select an algorithm

- Supervised Machine Learning (OLS,LASSO,XGboost, ANN, CNN, RNN, Transformer)
- Unsupervised Machine Learning (KNN,PCA,Autoencoder)
- Reinforcement Learning (DQN,PPO)
- The model should fit your data and business need

Model Evaluation and Optimization

- **Evaluate performance**: Test the model on the validation set.
- **Optimize hyperparameters**: Tune parameters to improve performance.
- **Address overfitting**: Use techniques like regularization or cross-validation.

Model Deployment

- **Present your model**: Present your model to the business partners or seminars for academics like a salesman
- **Deploy to production**: Integrate the model into a real-world system.
- **Monitor performance**: Track the model's performance in production.
- **Handle feedback**: Use feedback to improve the model.

Continuous Improvement

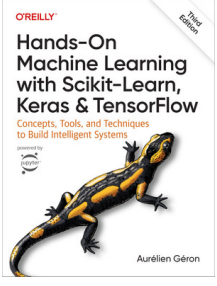
- **Periodic evaluation**: Regularly assess the model's accuracy.
- **Failure analysis**: Identify and address weaknesses.
- **Update and retrain**: Keep the model relevant with new data.

Conclusion

- Each stage is crucial for building effective models.
- DATA is huge important for the modeling
- Continuous improvement ensures long-term success.
- Don't forget to show your young talents to the audiences.

Textbook

- Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow
- Popular training textbooks for financial institutes, quants, or universities
- Pros: real practices
- Cons: lack math (The Elements of Statistical Learning for math)



Codes

- Codes: https://github.com/ageron/handson-ml3/tree/main
- Git Download: git clone https://github.com/ageron/handson-ml3.git

Read the textbook & run the codes

- Efficiently reading textbook and running codes are crucial for our modeling study
- Please follow me read the textbook and run the codes

Reference

1. Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow (3rd edition)
2. Andrew Ng's Coursera
3. Kaggle
4. Wikipedia
5. ChatGPT
6. DeepSeek