Why is Machine Learning in finance so hard?

Hardik Patel
Machine Learning Engineer, qplum

www.qplum.co/events

Changing Data Distributions

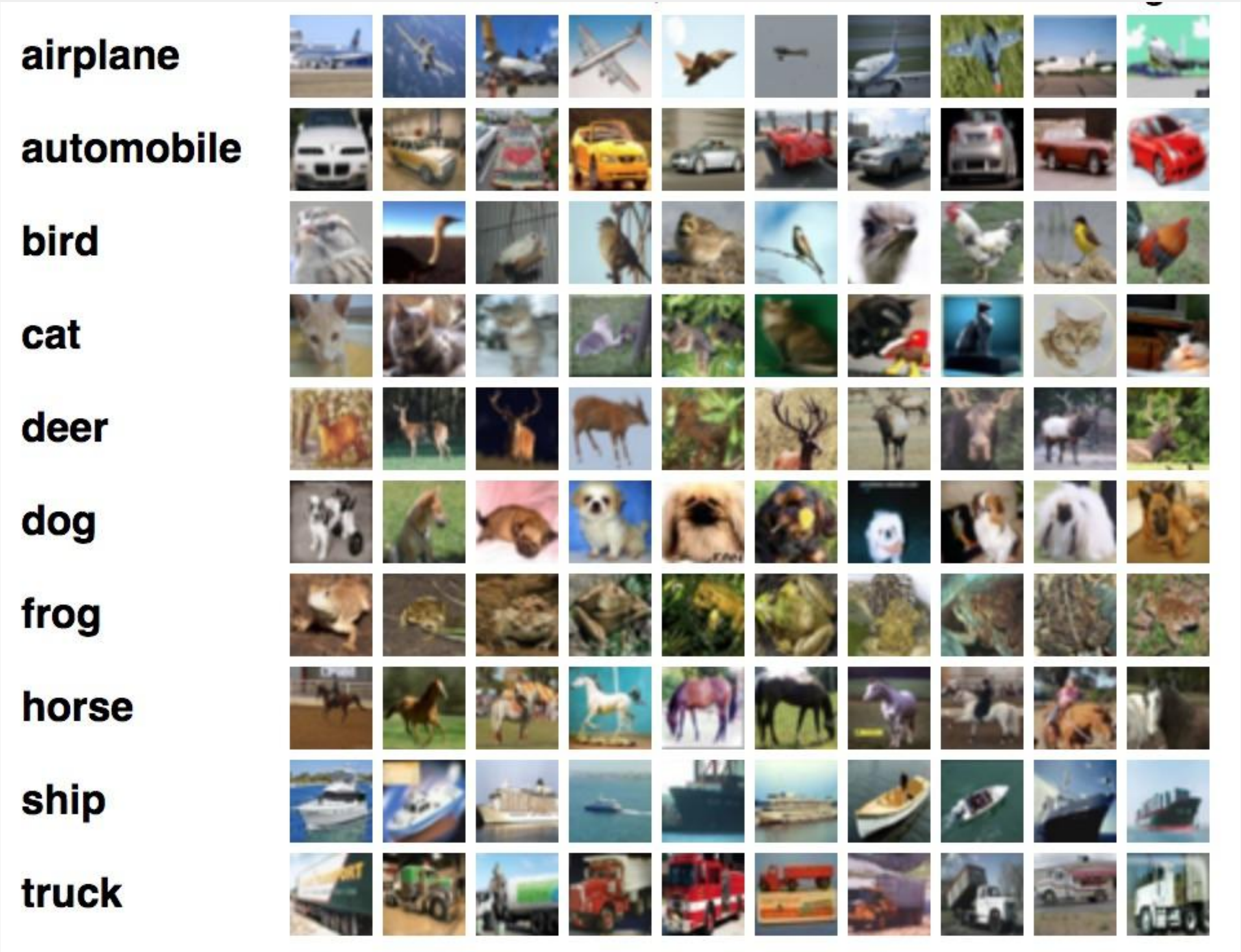
How does a typical quant pipeline look like?

- Start with a bunch of alphas
- Go long on the top 10
- Go short on the bottom 10

Source: www.hardikp.com

But, how does Machine Learning come into the picture?

CIFAR10 is an image classification dataset



Source: Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, The CIFAR-10 dataset, www.cs.toronto.edu/~kriz/cifar.html

"Understanding deep learning requires rethinking generalization" paper from Google Brain

Randomization tests. At the heart of our methodology is a variant of the well-known randomization test from non-parametric statistics (Edgington & Onghena, 2007). In a first set of experiments, we train several standard architectures on a copy of the data where the true labels were replaced by random labels. Our central finding can be summarized as:

Deep neural networks easily fit random labels.

More precisely, when trained on a completely random labeling of the true data, neural networks achieve 0 training error. The test error, of course, is no better than random chance as there is no correlation between the training labels and the test labels. In other words, by randomizing labels alone we can force the generalization error of a model to jump up considerably without changing the model, its size, hyperparameters, or the optimizer. We establish this fact for several different standard architectures trained on the CIFAR10 and ImageNet classification benchmarks. While simple to state, this observation has profound implications from a statistical learning perspective:

1. The effective capacity of neural networks is sufficient for memorizing the entire data set.
2. Even optimization on random labels remains easy. In fact, training time increases only by a small constant factor compared with training on the true labels.
3. Randomizing labels is solely a data transformation, leaving all other properties of the learning problem unchanged.

Is there a way to solve this generalization problem?

- Walk-forward Optimization

Low Predictive Power

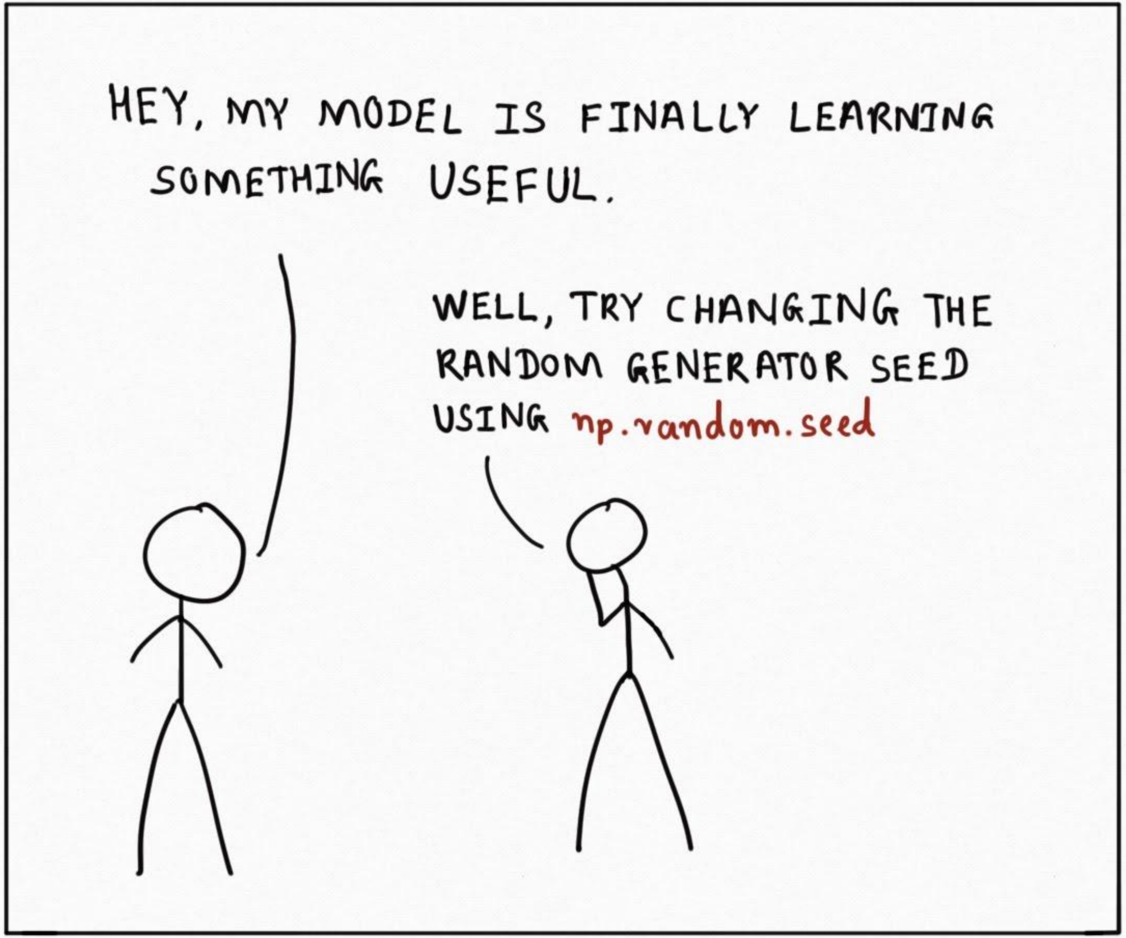
- Low accuracies compared to other domains.
- POMDP nature of the problem puts an implicit limit on the extent of what can be predicted.

So, what can we do here?

- Recognize that this is a different domain and adjust your expectations.
- While general prediction problems like return prediction are very hard, niche problems are easier to deal with.
- Focus on detecting regimes.

Low signal to noise ratio

If you see a pattern in the dataset, it's more likely to be noise than signal



Source: www.hardikp.com

Model interpretability

- There is almost always a chance of overfitting if you aren't able to answer touch interpretability questions.
- The challenge is to find the source of the model signal.

Model accuracy is not correlated with **utility function**

Model accuracy improvement might not lead to better portfolio returns

- Evidence for this can be found at all scales in financial markets: from high-frequency trading to long-term investing.

Strategy	MSE	CAR	Sharpe Ratio
S&P 500	n/a	4.5%	0.19
Market Avg.	n/a	7.7%	0.29
Price-LSTM	n/a	11.3%	0.60
QFM	0.62	14.4%	0.55
LFM-Linear	0.53	15.9%	0.63
LFM-MLP	0.47	17.1%	0.68
LFM-RNN	0.47	16.7%	0.67

(a) Out-of-sample performance for the 2000-2014 time period. All factor models use EBIT/EV. QFM uses current EBIT while our proposed LFMs use predicted EBIT. Price-LSTM is trained to predict price directly.

Source: Alberg, John & Lipton, Zachary C. Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals

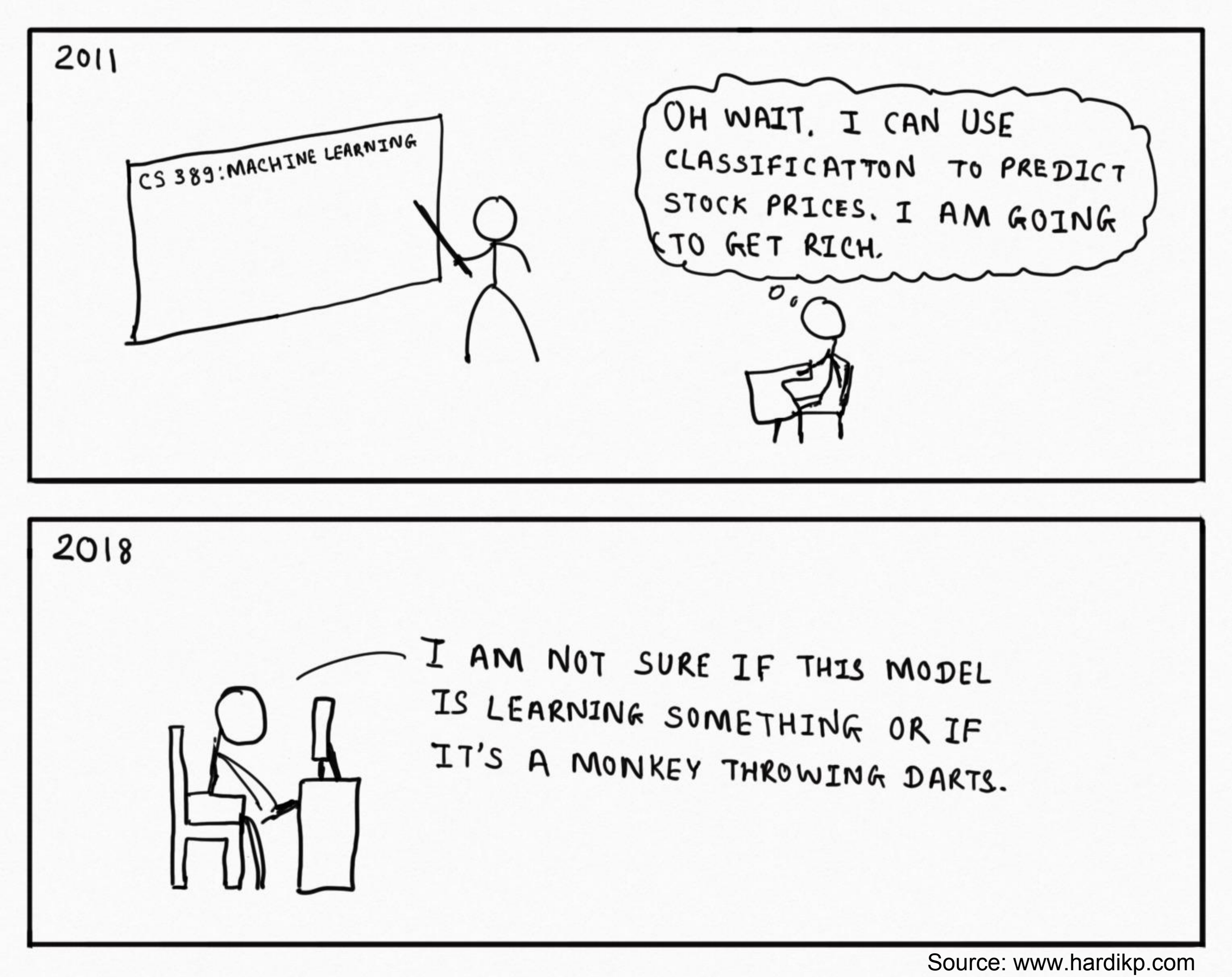
Wait, why can't we directly optimize the utility function?

1. Write down the utility function in terms of the raw data.
2. Use quadratic solvers or gradient descent to find the weights.

- This is a hard problem, often leading to simplification of the utility function.
- The optimization process is unstable and unreliable.

Reinforcement Learning is another option to directly optimize the utility function

- Doesn't suffer from oversimplification issues.
- RL state can be made sufficiently complex.
- But the state doesn't have enough information to guide the agent in the right direction.
- Lack of enough predictive information often causes the agents to wander in random directions.



Source: www.hardikp.com

Questions?

contact@qplum.co