ENSEMBLING

Rick Fontenot
SMU DS7335 Machine Learning II
rickfontenot@gmail.com

Rapid Rise in Research

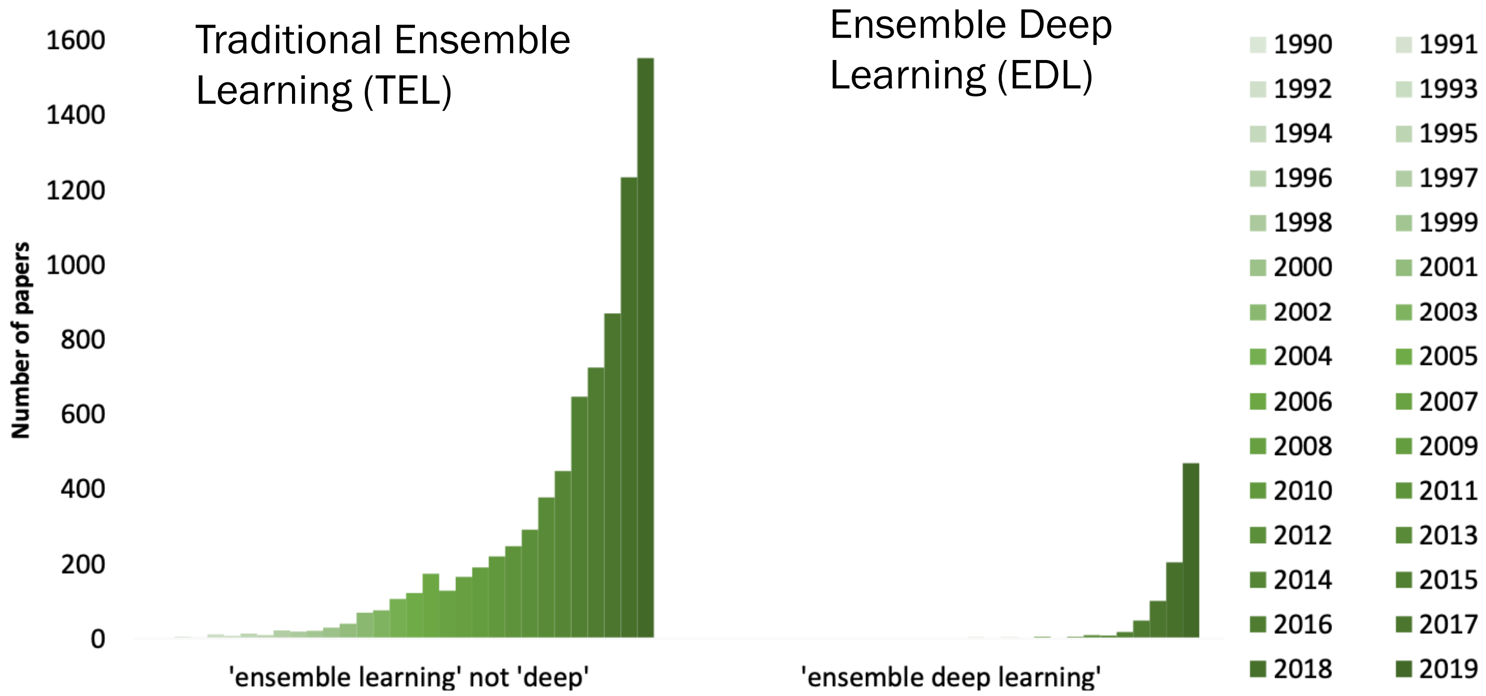


Figure 2. The comparison between the number of papers published in the core set of Web of Science from 1990 to 2019 for the topic of 'traditional ensemble learning' and the topic of 'ensemble deep learning'.

Yongquan Yang et. al "A Survey on Ensemble Learning under the Era of Deep Learning" 2021

Popularity in Competitions

The leaderboard of top score for a neural network challenge as of November 2020. The best single model was in position 7 with an EM score of 89.551.

Slide From Dr. Slater's Quantifying the World Course

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748

Counterpoint & Considerations

"More often than not, winners of Kaggle competitions have used ensemble methods…while they are a popular choice for ML competitions, they are not used in production in the "real world" quite as often as we might expect" *--Philip Tannor, CEO of Deepchecks (uses Explainable AI to check and monitor ML models)*

When you shouldn't use Ensemble Learning:

- Can't afford extra overhead in training and inference time
- If your model must operate in real-time it needs to be lean to reduce latency.
- Additional training time if you need to regularly retrain to avoid model drift.
- Lack of explain-ability

Tanoor al "When You Shouldn't Use Ensemble Learning" 2021
https://deepchecks.com/when-you-shouldnt-use-ensemble-learning/

Basic Approach

Generally, an ensemble is constructed in two steps, i.e., generating the base learners, and then combining them. To get a good ensemble, it is generally believed that the base learners should be as *accurate* as possible, and as *diverse* as possible.

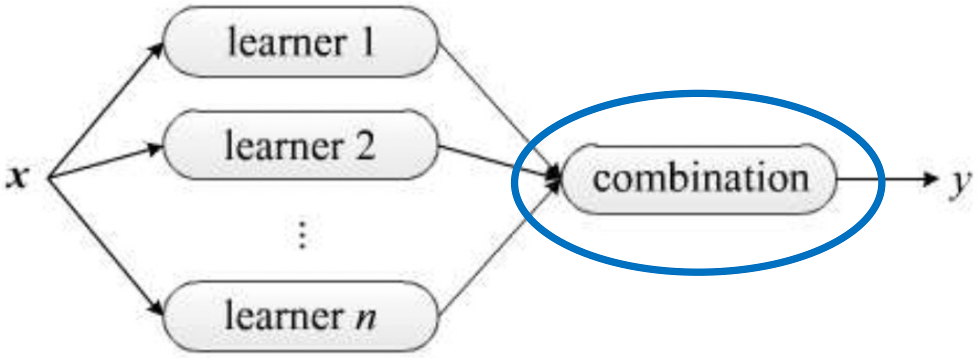


FIGURE 1.9: A common ensemble architecture.

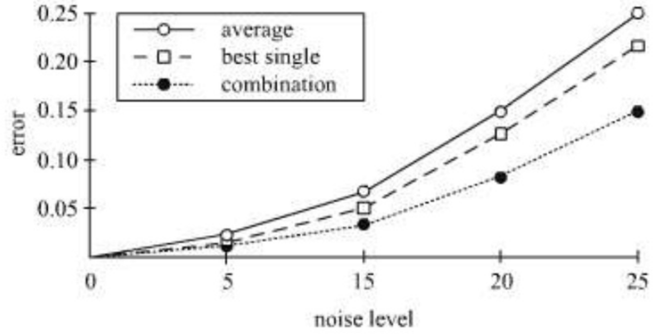


FIGURE 1.10: A simplified illustration of Hansen and Salamon [1990]'s observation: Ensemble is often better than the best single.

Zhi-Hua Zhou "Ensemble Methods: Foundations & Algorithms" 2012

Combination Methods

Numerical predictions:

- Simple Averaging
- Weighted averaging

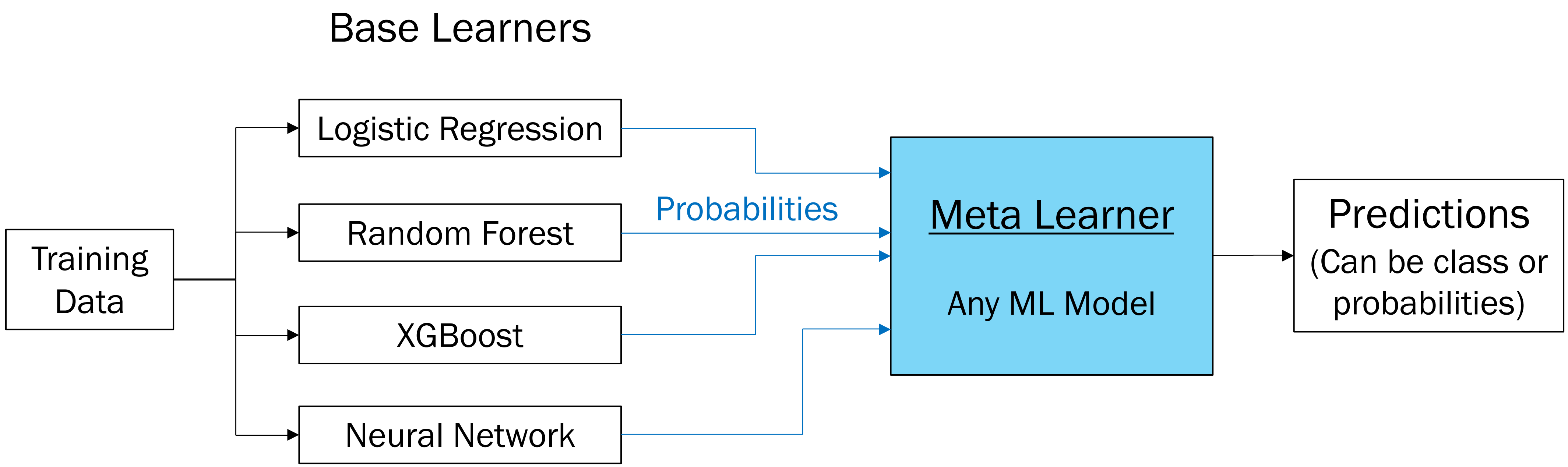
Classification:

- Majority voting – if a class receives more than 50% votes between base learners that becomes the ensemble prediction. If no class receive 50% then there is no ensemble prediction and you can default to a class or choose one base learner's results as default.
- Plurality voting – simply choose the class with the most votes between all base learners
- Weighted voting
- Soft voting – rather than using hard classifications from each base learner, average their prediction probabilities. This can be weighted averaging similar to above

Meta Learners:

- Combination determine by more complex models such as regression, random forest, neural networks or an appropriate type
- Can be used for Numerical or Classification predictions

Meta-Learners



Assumptions: Out-of-fold predictions

On one hand, stacking is a general framework which can be viewed as a generalization of many ensemble methods. On the other hand, it can be viewed as a specific combination method which combines by learning, and this is the reason why we introduce Stacking in this chapter.

In the training phase of stacking, a new data set needs to be generated from the first-level classifiers. If the exact data that are used to train the first-level learner are also used to generate the new data set for training the second-level learner, there will be a high risk of overfitting. Hence, it is suggested that the instances used for generating the new data set are excluded from the training examples for the first-level learners, and a cross-validation or leave-one-out procedure is often recommended.

sulting learner h' is a function of $(z_1, \ldots, z_T)$ for y. After generating the new data set, generally, the final first-level learners are re-generated by training on the whole training data.

Zhi-Hua Zhou "Ensemble Methods: Foundations & Algorithms" 2012

Assumptions: Model Diversity

Ensemble diversity, that is, the difference among the individual learners, is a fundamental issue in ensemble methods.

Intuitively it is easy to understand that to gain from combination, the individual learners must be different, and otherwise there would be no performance improvement if identical individual learners were combined.

So, it is desired that the individual learners should be *accurate and diverse*. Combining only accurate learners is often worse than combining some accurate ones together with some relatively weak ones, since complementarity is more important than pure accuracy. Ultimately, the success of ensemble learning lies in achieving a good tradeoff between the individual performance and diversity.

Unfortunately, though diversity is crucial, we still do not have a clear understanding of diversity; for example, currently there is no well-accepted formal definition of diversity. There is no doubt that understanding diversity is the holy grail in the field of ensemble learning.

Zhi-Hua Zhou "Ensemble Methods: Foundations & Algorithms" 2012

Assumptions: Pruning

What Is Ensemble Pruning

Given a set of trained individual learners, rather than combining all of them, **ensemble pruning** tries to select a subset of individual learners to comprise the ensemble.

An apparent advantage of ensemble pruning is to obtain ensembles with smaller sizes; this reduces the storage resources required for storing the ensembles and the computational resources required for calculating outputs of individual learners, and thus improves efficiency. There is another benefit, that is, the generalization performance of the pruned ensemble may be even better than the ensemble consisting of all the given individual learners.

Zhi-Hua Zhou "Ensemble Methods: Foundations & Algorithms" 2012

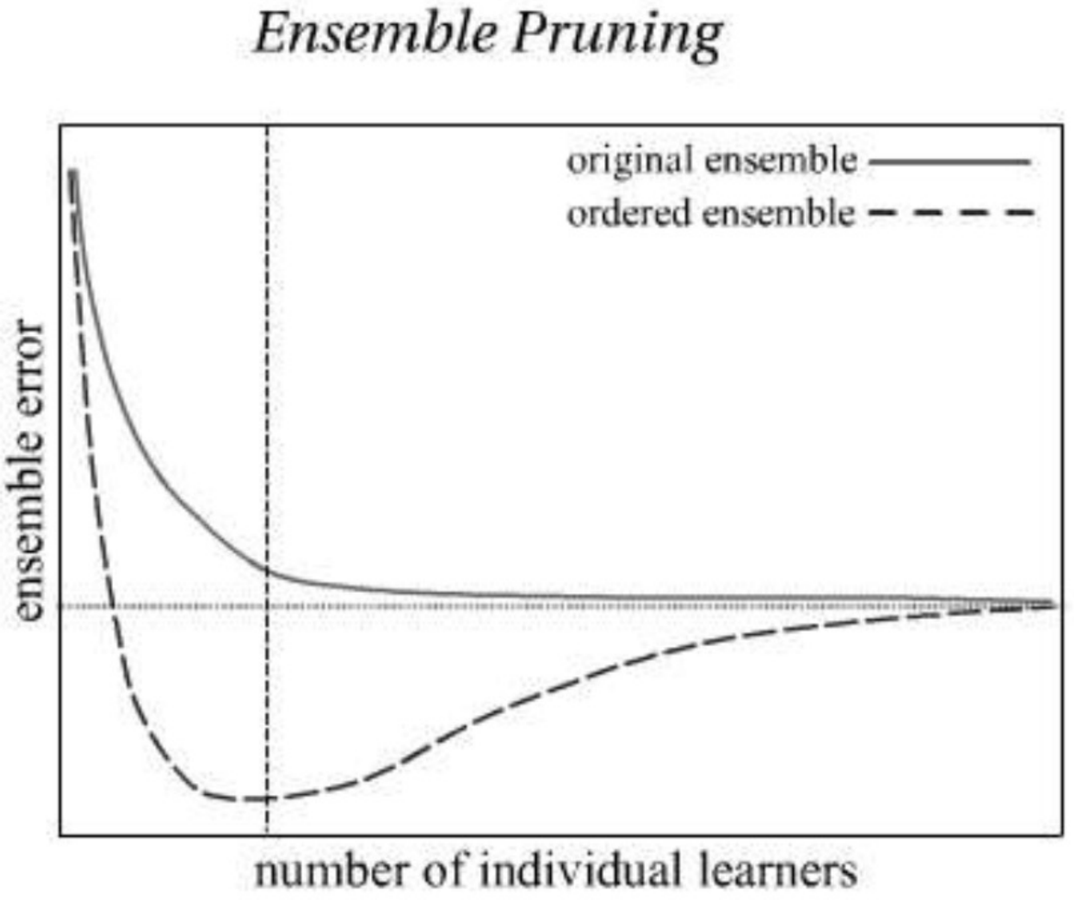
Assumptions: Pruning

Order of base learners in pruning process can be ranked:

- by lowest validation error
- by highest diversity pairings of base learners

Forward Selection

If the order is random, you may not find the optimal pruned mixture



Zhi-Hua Zhou "Ensemble Methods: Foundations & Algorithms" 2012

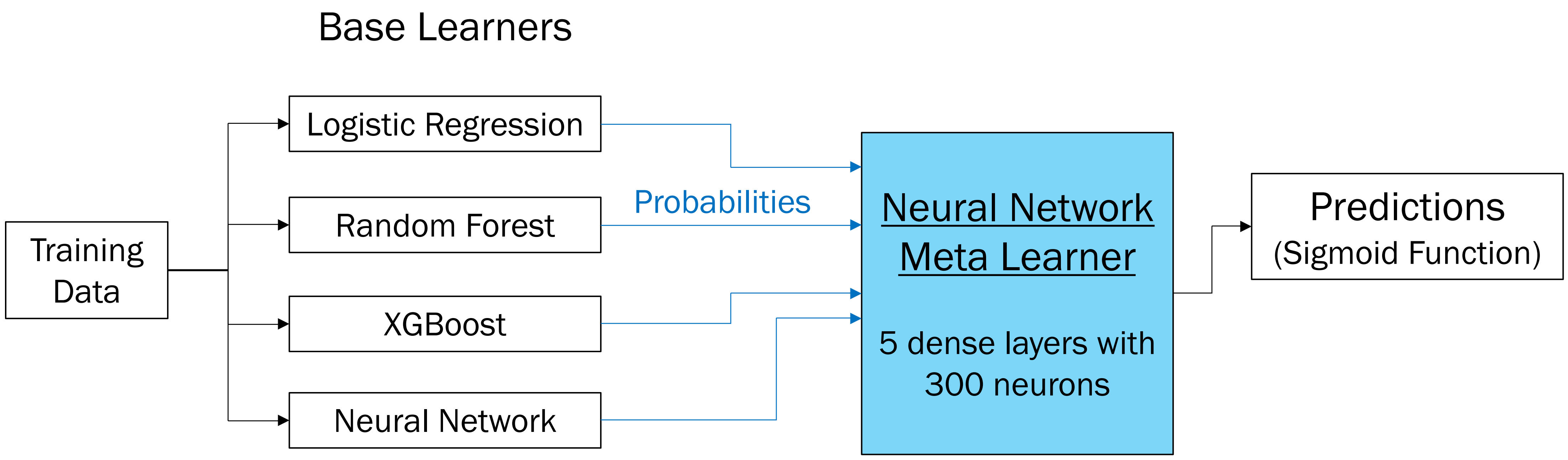
DEMONSTRATION

Case study to explore effects of ensemble assumptions & compare results and modeling time

Background info on case for demo

- Anonymized data set with 45 numeric features and 5 string features
- Target is binary and also anonymized, client wants predictions to decide what actions to take on each transaction
- Incorrect predictions cost the client money on each transaction, $100 for a false positive and $200 for a false negative
- Cost function was set up to sum all these costs on prediction set then divide by total predictions to normalize to Average Money Lost per transaction
- The client has approximately 160,000 transactions annually, so the annual value of savings can be estimated for each model
- Goal is to minimize money loss (not maximize accuracy)

Neural Network Meta-Learner



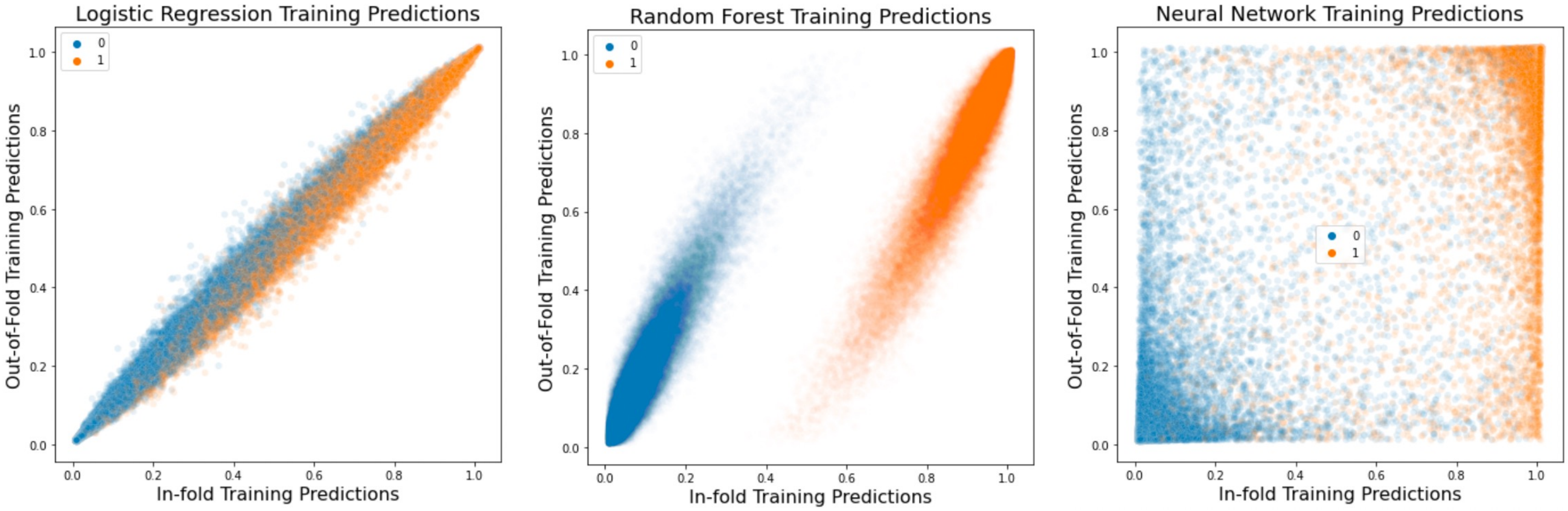
Base Learner Models

Model	Training Time	Val Predict Time	Train OOF Predict Time	Test Accuracy	Test Cost
Neural Network	19.7	0	82.4	97.5	$ 4.13
XGB Classifier	0.8	0	4.7	93.4	$ 9.10
Random Forest	1.3	0	4.8	91.7	$11.36
Logistic Regression	0.2	0	0.9	87.3	$22.89

- Neural Network (NN) is best individual model
- NN train time is longer, creating out-of-fold predictions on training set much longer
- Time to create validation set predictions is negligible after models are trained
- Will validation set have large enough sample size for meta-learner ensemble?

In-fold vs. out-of-fold predictions

- Predictions are relatively similar in linear model, but more significantly different in more complex models prone to overfitting
- With a 0.5 cutoff, the 0/1 classifications have similar error rates but the probabilities are what feed as inputs to meta-learner so the differences can be significant when overfit



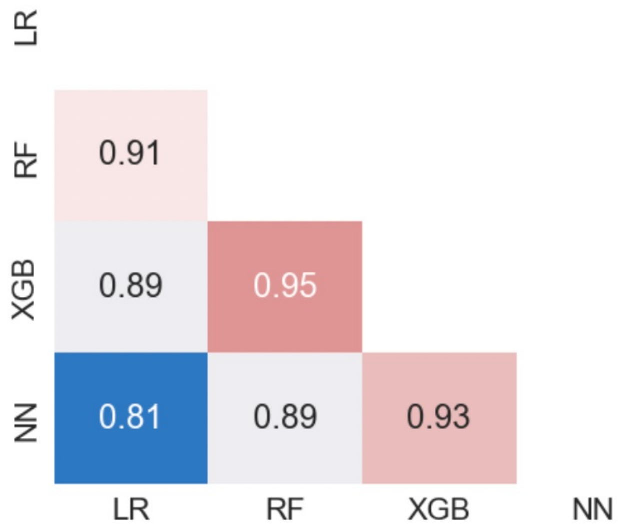
Comparison of Ensembles trained from in-fold vs. out-of-fold predictions

Ensemble Training Data	Sample Size	Test Accuracy	Test Cost
Individual NN Base Leaner without ensembling	126713	97.45	$ 4.13
NN Meta-learner with In-fold probabilities from training set	126713	92.56	$ 11.76
NN Meta-learner with Out-of-fold probabilities from training set	126713	97.47	$ 3.94
NN Meta-learner with Out-of-fold probabilities from validation set	15839	97.44	$ 3.94

- All 4 base learners included in a neural network meta-learner ensemble
- Using In-fold probabilities from base learner performed significantly worse than not ensembling at all (predictions were overfit)
- Out-of-fold predictions on smaller validation set performed as well as larger training set
- For this case study, additional time to create out-of-fold predictions on training set not worth the time, but in other cases with smaller sample sizes it may be advisable

Pruning Exercises by Ranking Order

Correlation Matrix for OOF predictions



Pruning Ex. 1

Order by Best Loss:
Neural Network
XGB Classifier
Random Forest
Logistic Regression

Pruning Ex. 2

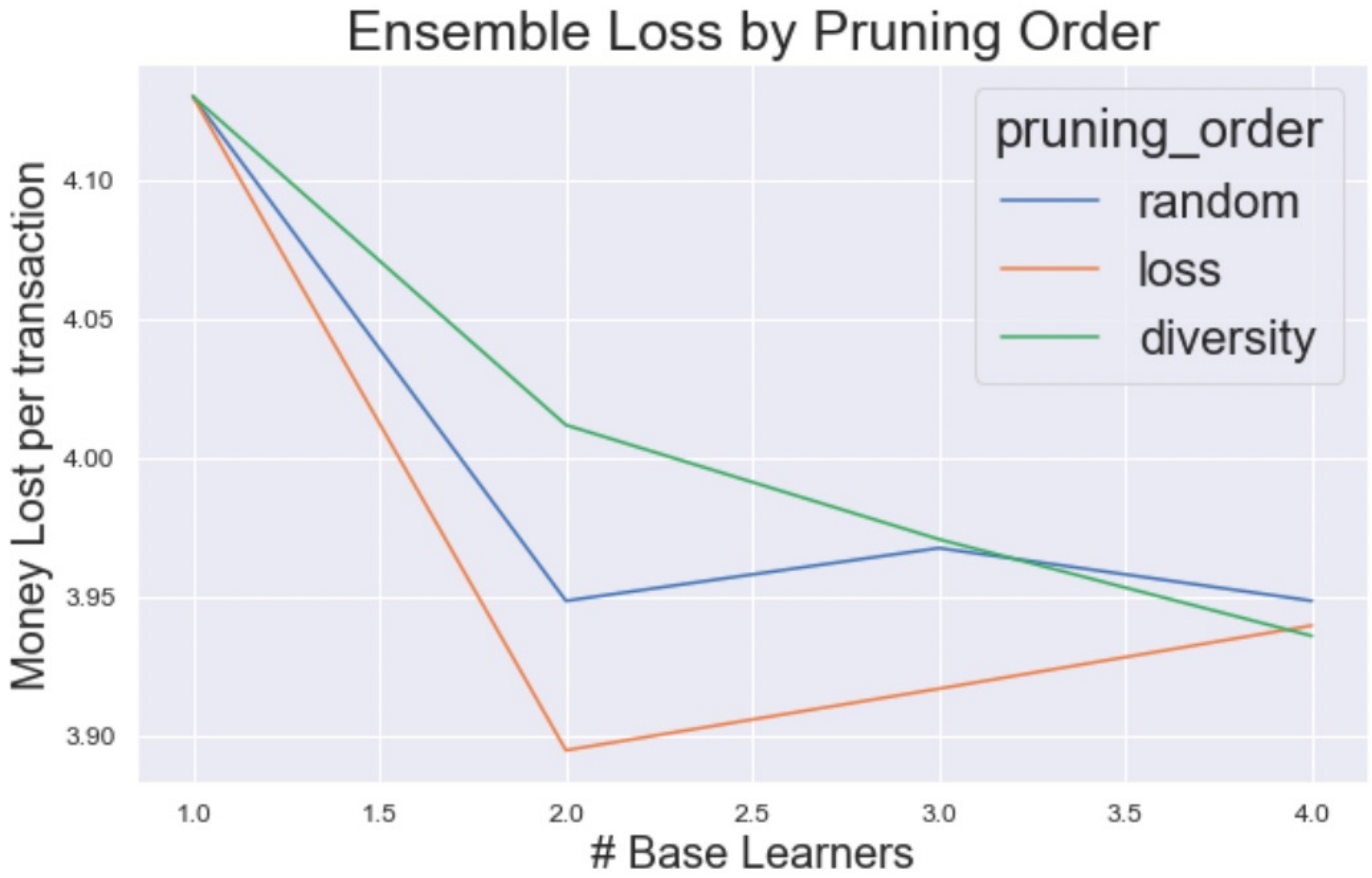
Order by Diversity:
Neural Network
Logistic Regression
Random Forest
XBG Classifier

Pruning Ex. 3

Order Randomly:
Neural Network
Random Forest
Logistic Regression
XBG Classifier

Pruning a Neural-Network Meta-Learner

- Ranking by Loss produced the best ensemble with just two learners
- Reducing base learners to just NN & XGB lowers loss and saves time vs. using all 4 base learners



Comparison of ensemble models

- Best Ensemble saves $38,400 annually with less than 3 additional minutes of training time compared to best individual model
- Without exploring out-of-fold sample size and without pruning, ensemble training time would have taken 5X as long and saved less money than the final pruned ensemble

Model	**Modeling Time**	**Test Accuracy**	**Money Lost per transaction**	**Money Lost Annually**
Individual Neural Network	19.7	97.5	$ 4.13	$ 660,800
NN Meta-Learner with all 4 base learners, oof from training set	116.8	97.5	$ 3.94	$ 630,400
NN Meta-Learner with all 4 base learners, oof from validation set	24	97.4	$ 3.94	$ 630,400
NN Meta-Learner with NN+XGB base learners, oof from validation set	22.5	97.5	$ 3.89	$ 622,400

*** Modeling Time = (base learner training times) + oof prediction time + meta-learner training**

Re-cap so far

- Out-of-fold predictions are critical
- “Good” models with low loss are critical, but model diversity is important too
- Pruning ensembles produces better results than just ensembling all the base learners available, and it saves time by reducing number of models
- Training time, compute resources and time to make out-of-fold predictions for meta-learner inputs all create challenges for Deep Ensemble Learning
- **So what’s happening in the realm of Deep Ensemble Research?**

Recent Deep Ensemble research

- Suk HI, Lee SW, Shen D; Alzheimer's Disease Neuroimaging Initiative
- Challenge in brain imaging analysis is the high dimensionality of data, but with a small number of samples available
- Due to interpretable model requirements, sparsity-inducing penalization is considered as one of the key techniques for feature selection in medical problems.
- They built multiple sparse regression models with different values of a regularization control parameter
- CNN as ensemble by taking the predictions from the multiple regression models as input for final clinical decision making
- Higher sensitivity score means the lower the chance of mis-diagnosing patients
- This ensemble method improved sensitivity by 4.36% to 7.77% compared to previous modeling methods on the same subject

Suk HI, Lee SW, Shen D; Alzheimer's Disease Neuroimaging Initiative. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal. 2017 Apr;37:101-113. doi: 10.1016/j.media.2017.01.008. Epub 2017 Jan 24. PMID: 28167394; PMCID: PMC5808465.

Recent Deep Ensemble research

- W. Liu, M. Zhang, Z. Luo and Y. Cai, "An Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors,"
- Balanced sampling data augmentation strategy to increase the number of samples of rare classes in the original dataset
- Multiple Convolutional neural network models are trained with different residual learning frameworks (initialization pretrained on ImageNet for all)
- Outputs of CNN models combined by maximum voting policy

Model	Mean Recall	Precision	Mean Precision	Cohen Kappa Score
ResNet-50	0.8244	0.9586	0.8684	0.9354
ResNet-50-BS	0.8639	0.9610	0.8648	0.9392
ResNet-101	0.8713	0.9691	0.8713	0.9520
ResNet-101-BS	0.8841	0.9705	0.8956	0.9540
ResNet-152	0.8773	0.9698	0.8978	0.9531
ResNet-152-BS	**0.8882**	0.9713	0.8929	0.9553
DCEM(ours)	0.8708	0.9723	0.9106	0.9568
DCEM-BS(ours)	0.8844	**0.9776**	**0.9201**	**0.9651**

Best individual CNN

Ensemble

W. Liu, M. Zhang, Z. Luo and Y. Cai, "An Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors," in IEEE Access, vol. 5, pp. 24417-24425, 2017, doi: 10.1109/ACCESS.2017.2766203.

Recent Deep Ensemble research

- Huang G, et al: Snapshot Ensembles: Train 1, Get M for Free
- Ensembles of neural networks more robust than individual networks but training multiple deep networks for model averaging is computationally expensive
- In gradient descent, number of possible local minima grows exponentially with number of parameters. Two identical architectures optimized with different initializations or minibatch orderings will converge to different solutions.
- Although different local minima often have very similar error rates, the corresponding neural networks tend to make different mistakes. This diversity can be exploited through ensembling, in which multiple neural networks are trained from different initializations and then combined with majority voting or averaging
- Snapshot Ensembling saves models at each local minimum then continues training
- Ensemble of snapshot models yielded lower error rates than single models at no additional training cost, and compare favorably to traditional network ensembles

Huang G, Li Y, Pleiss G, et al (2017a) Snapshot Ensembles: Train 1, Get M for Free. In: International Conference on Learning Representations 2017

Snapshot Ensemble Results

- Notice that pruning matters here too, adding more snapshots to ensemble doesn't keep reducing error rate

Method	**Val. Error (%)**
Single model	24.01
Snapshot Ensemble ($M = 2$)	23.33
Snapshot Ensemble ($M = 3$)	23.96

Table 2: Top-1 error rates (%) on ImageNet validation set using ResNet-50 with varying number of cycles.

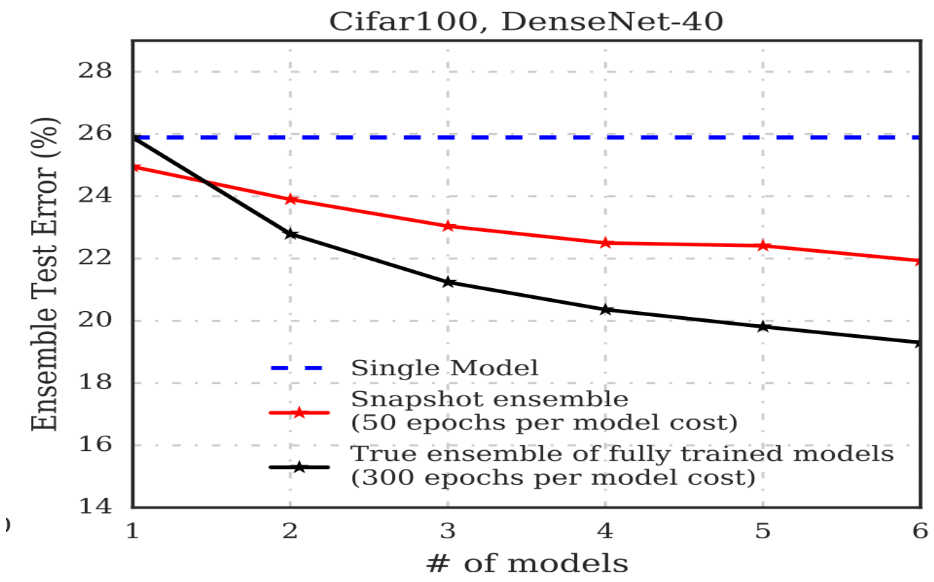
M	**Test Error (%)**
2	22.92
4	22.07
6	21.93
8	21.89
10	22.16

Table 3: Error rates of a DenseNet-40 Snapshot Ensemble on CIFAR-100, varying M—the number of models (cycles) used in the ensemble.

Huang G, Li Y, Pleiss G, et al (2017a) Snapshot Ensembles: Train 1, Get M for Free. In: International Conference on Learning Representations 2017

Snapshot Ensemble Results

- Snapshot Ensemble didn't produce as low error as true ensemble of multiple fully trained models
- But remember that red line is #snapshots in one model, so at the right end of chart it's training time is approximately 1/6th that of the fully trained true ensemble



Huang G, Li Y, Pleiss G, et al (2017a) Snapshot Ensembles: Train 1, Get M for Free. In: International Conference on Learning Representations 2017

References

A Survey on Ensemble Learning under the Era of Deep Learning:
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/2101.08387.pdf

A Comparative Study of non-deep learning, deep learning, and ensemble learning methods:
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/2203.05757.pdf

Ensemble Methods: Foundations & Algorithms (book By Zhi-Hua Zhou):
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/EMFA_compressed.pdf

Clustering ensembles of Neural Network Models:
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/1-s2.0-S0893608002001879-main.pdf

Alzheimer's Disease Neuroimaging Initiative:
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/1-s2.0-S1361841517300166-main.pdf

An Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors:
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/An_Ensemble_Deep_Learning_Method_for_Vehicle_Type_Classification_on_Visual_Traffic_Surveillance_Sensors.pdf

Snapshot Ensembles: Train 1, Get M for Free
https://github.com/rickfontenot/ML2/blob/main/ensemble/research/snapshot_ensembles_train_1_get.pdf

Appendix