Southern University of
Science and Technology

Big Data: New Tricks for Econometrics [Hal R. Varían, 2014]

Yixuan Liu
Supervisor: MA, Yifang

2020-12-15

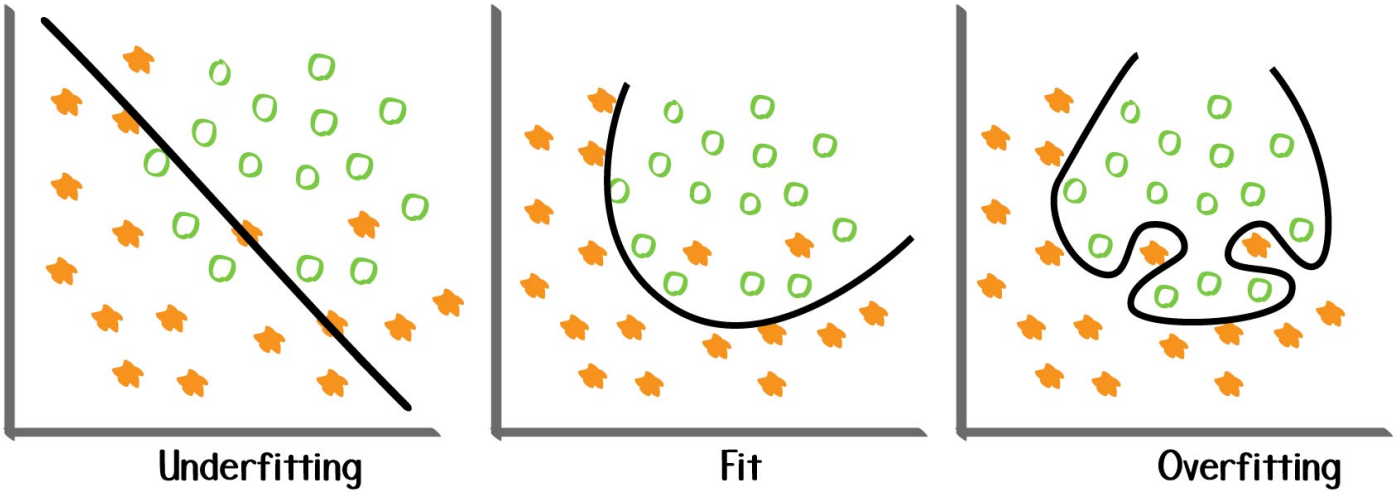
Department of Statistics and Data Science

Contents:

SUSTech

Overfitting problem

- **An overfitted model**: a statistical model that contains more parameters than can be justified by the data
- Cause:
  - The model is complex, with unneeded variables
  - Data has noise, i.e. outliers and errors
  - Size of data is small
- Consequences:
  - Poor performance on validation; costly; less portable



Overfitting remedies

- Mentioned by author:
  - Regularization
  - Dividing datasets
  - Cross-validation
  - Network-reduction (pre-pruning and post-pruning)
  - Ensembling (bagging, boosting)
- More to be covered:
  - Expansion of the training data
    1. Acquire more training data
    2. Add some random noise
    3. Produce data based on existing distribution
  - Remove features (feature selection)
  - Early stopping (when training iteratively)

Early stopping

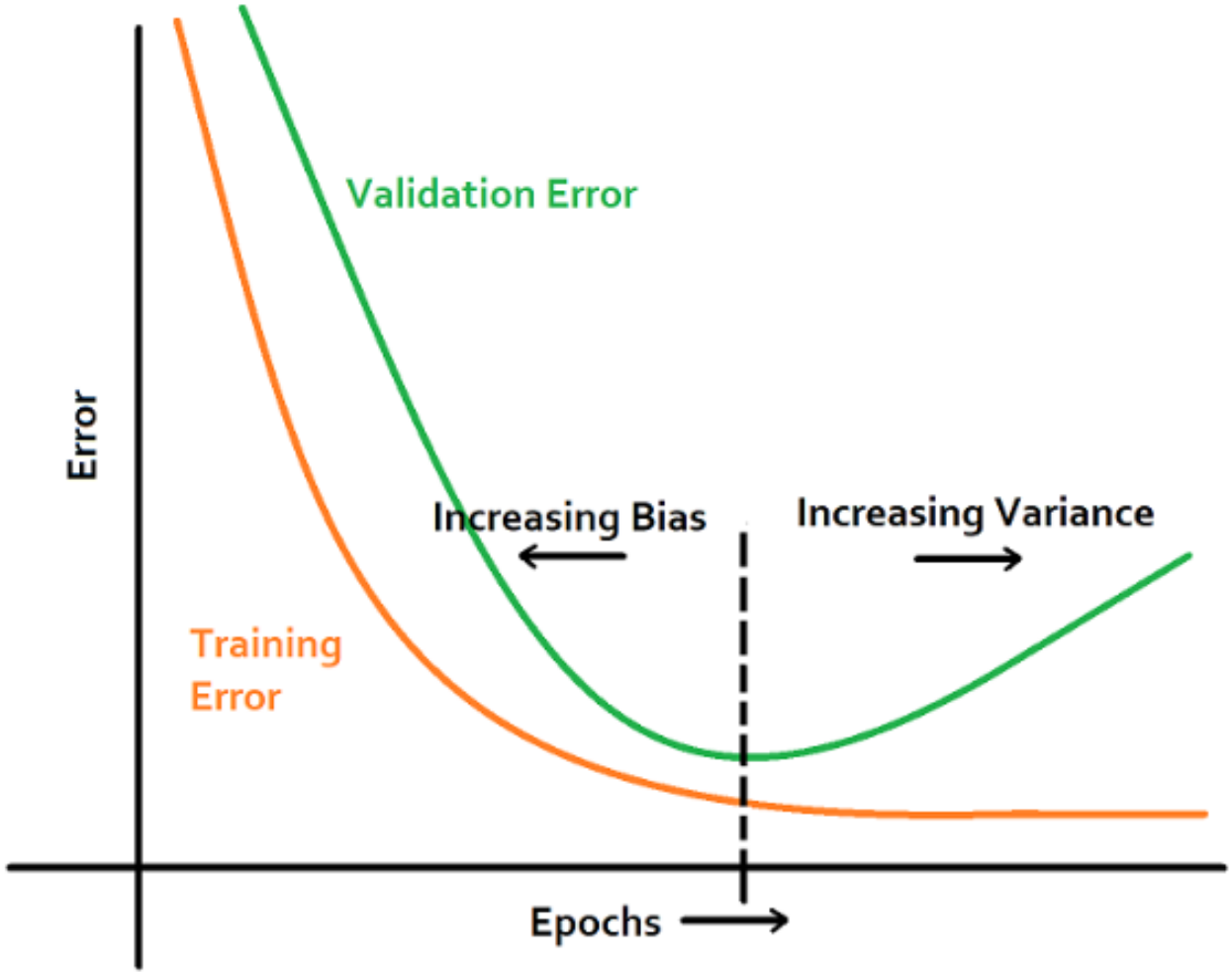
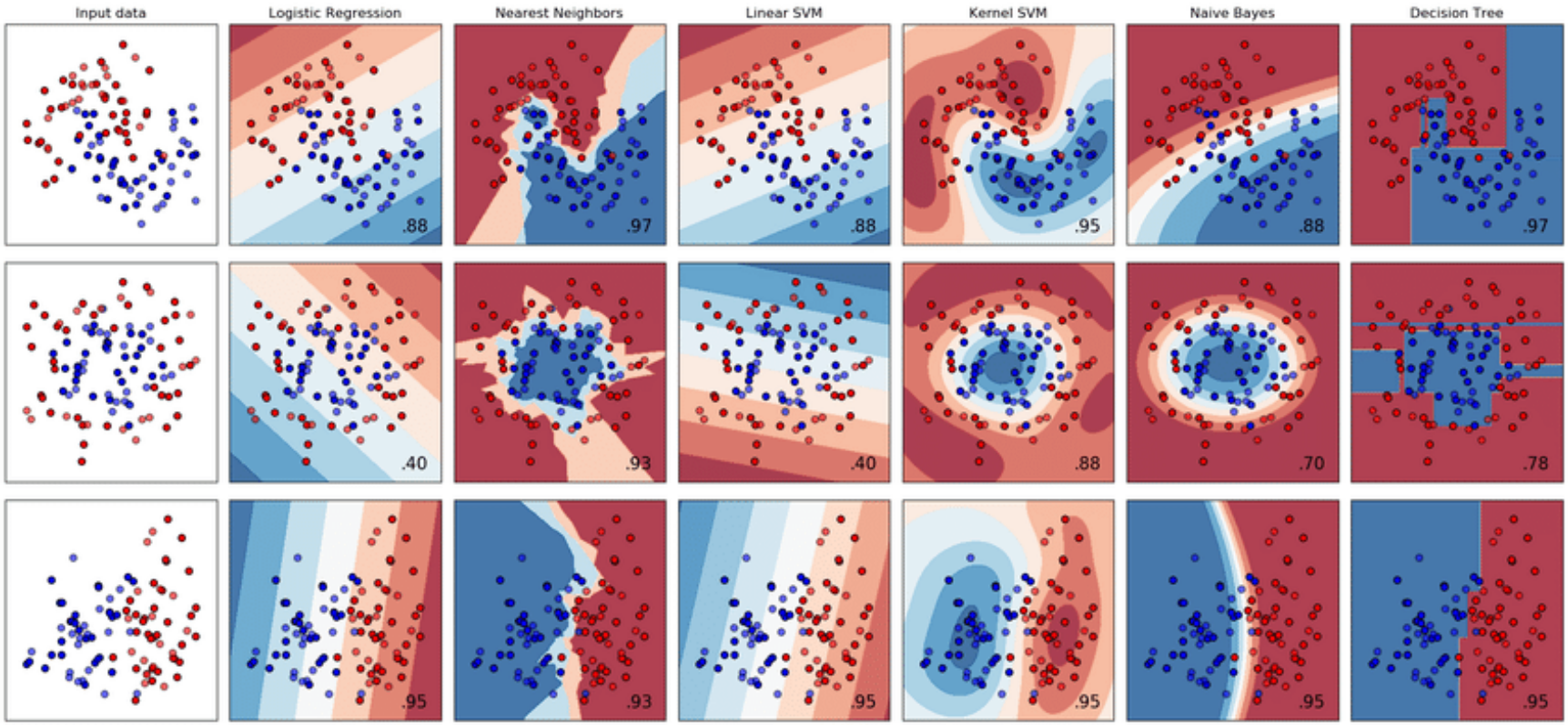


Figure: Validation error vs testing error

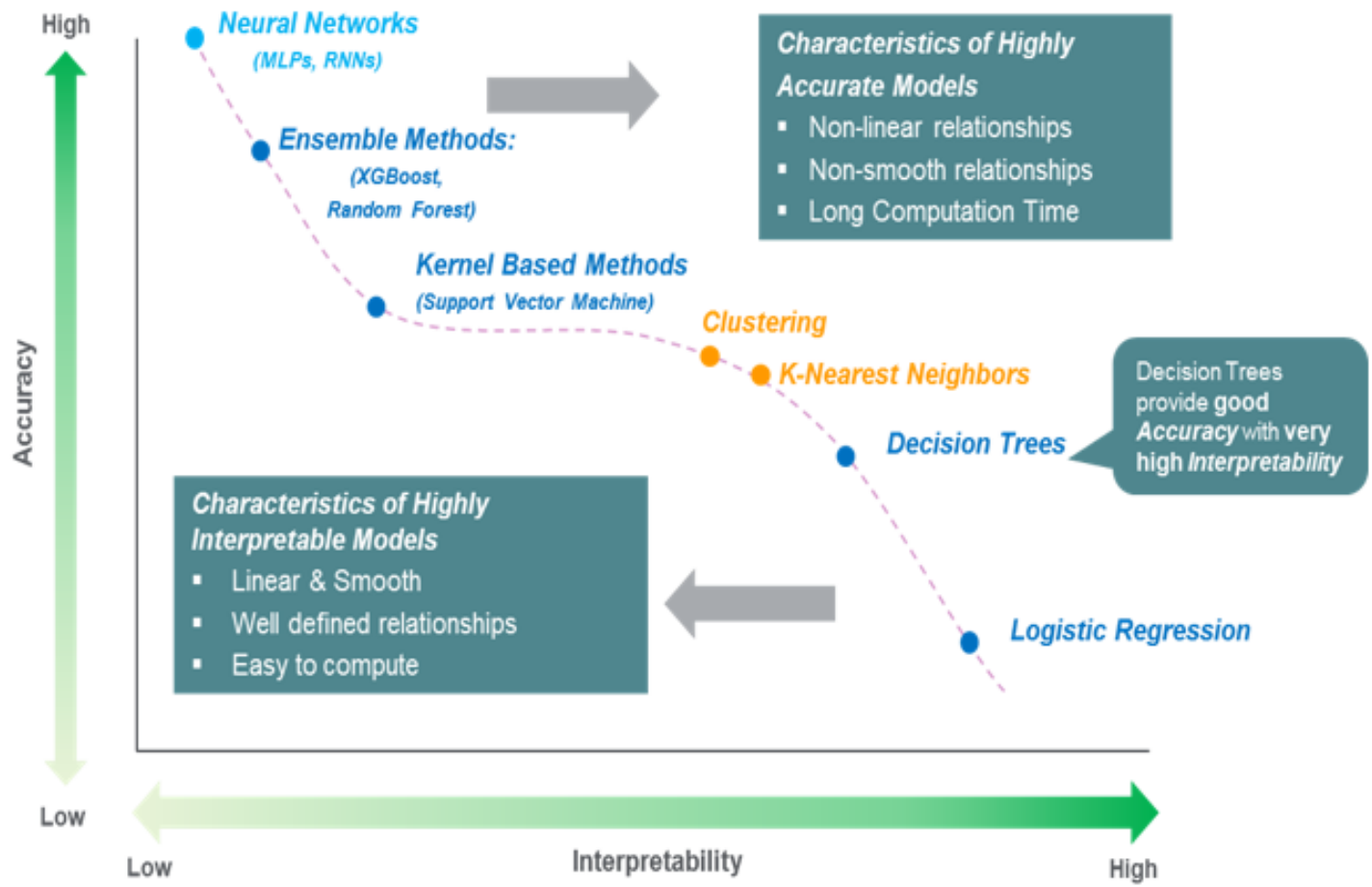
Classification

- Logistics regression:
  - **Pros**: better performance with small datasets; output could be interpreted as probability.
  - **Cons**: do not perform well on nonlinear data; apt to overfitting.
  - **Improvement**: imported regularization to avoid overfitting.
- Decision trees:
  - **Pros**: automatically select important attributes; strong interpretation.
  - **Cons**: fit poorly for small dataset; overfitting; results trends to majority class.
  - **Improvement**: balanced datasets with SMOTE (Synthetic Minority Oversampling Technique).

Classification region



Interpretability v.s. Accuracy



Unsupervised Learning

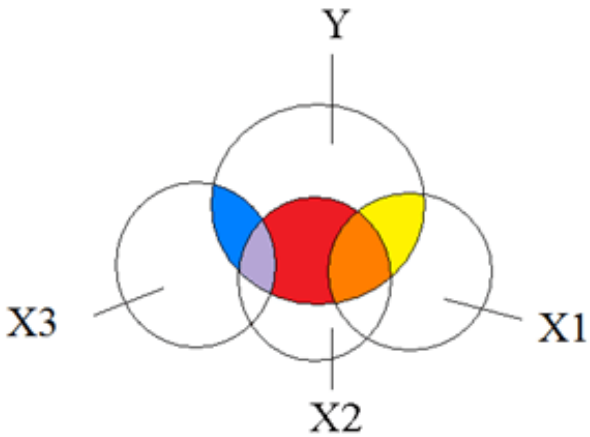
Supervised Learning

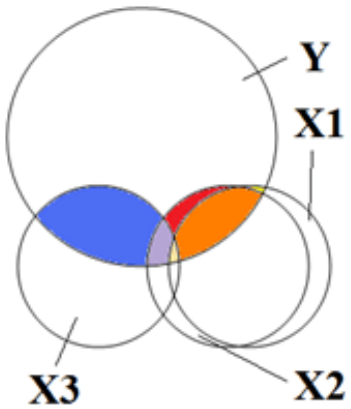
Both

SUSTech

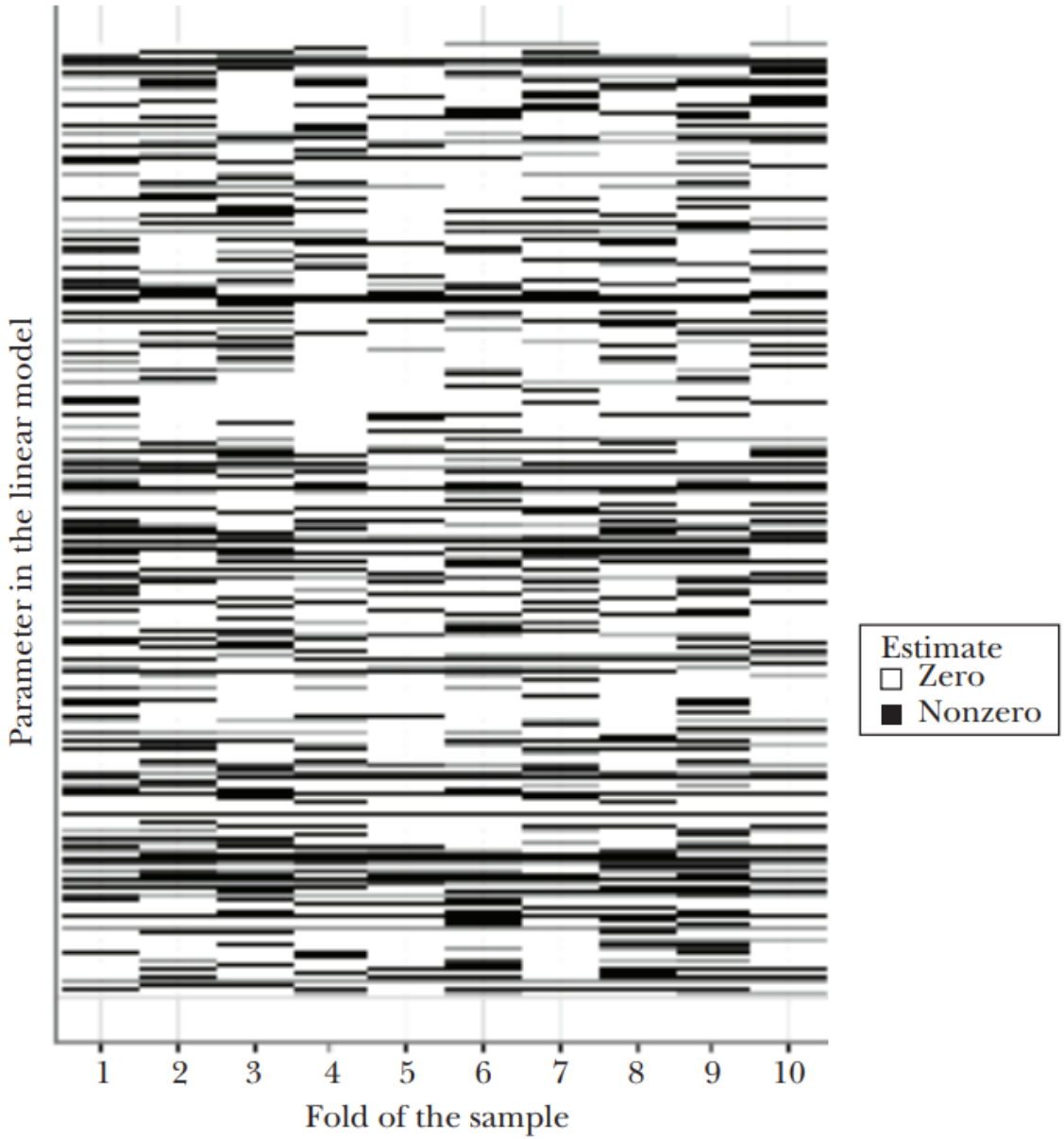
Multicollinearity

- Reason: variables are highly correlated
- Example: **mortgage data**, high correlation between race and denied mortgage insurance (dmi).
- Harm: increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model.
- Remedies: PCA, ridge regression, feature engineering



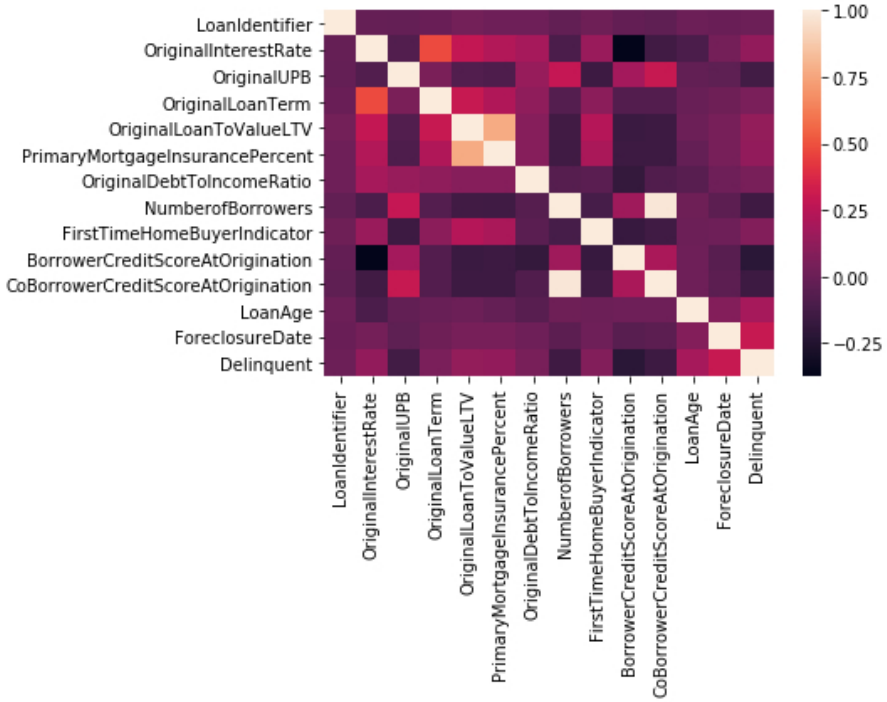


Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



Example of multicollinearity: mortgage delinquency

- PrimaryMortgageInsurancePercent and NumberOfBorrowers have been removed due to high correlation
- The ROC curve in particular seems indicate an average explanatory power for the model with an area under the curve (AUC) of 0.84.



Bayesian structural time series (BSTS)

- Pros:
  - Effectively prevent overfitting and spurious correlation
  - Useful for fat regression where attributes are more than observed values
  - Discover the causations with counterfactual prediction and observed data
  - In contrast to DID:
    1. Infer the temporal evolution of attributable impact
    2. Incorporate empirical priors on the parameters in a fully Bayesian treatment
    3. Flexibly accommodate multiple sources of variation (seasonality and etc.)
- Cons:
  - Relatively complicated mathematical underpinning

Causality and prediction

- Problem: number of machine learning algorithms could not depict causation
- Background: effect of advertising on sales, many confounding factors
- Unconventional design of control group: forecast visits *would have been* using BSTS, comparing the actual visits to counterfactual visits gives an estimate of causal effect of advertising

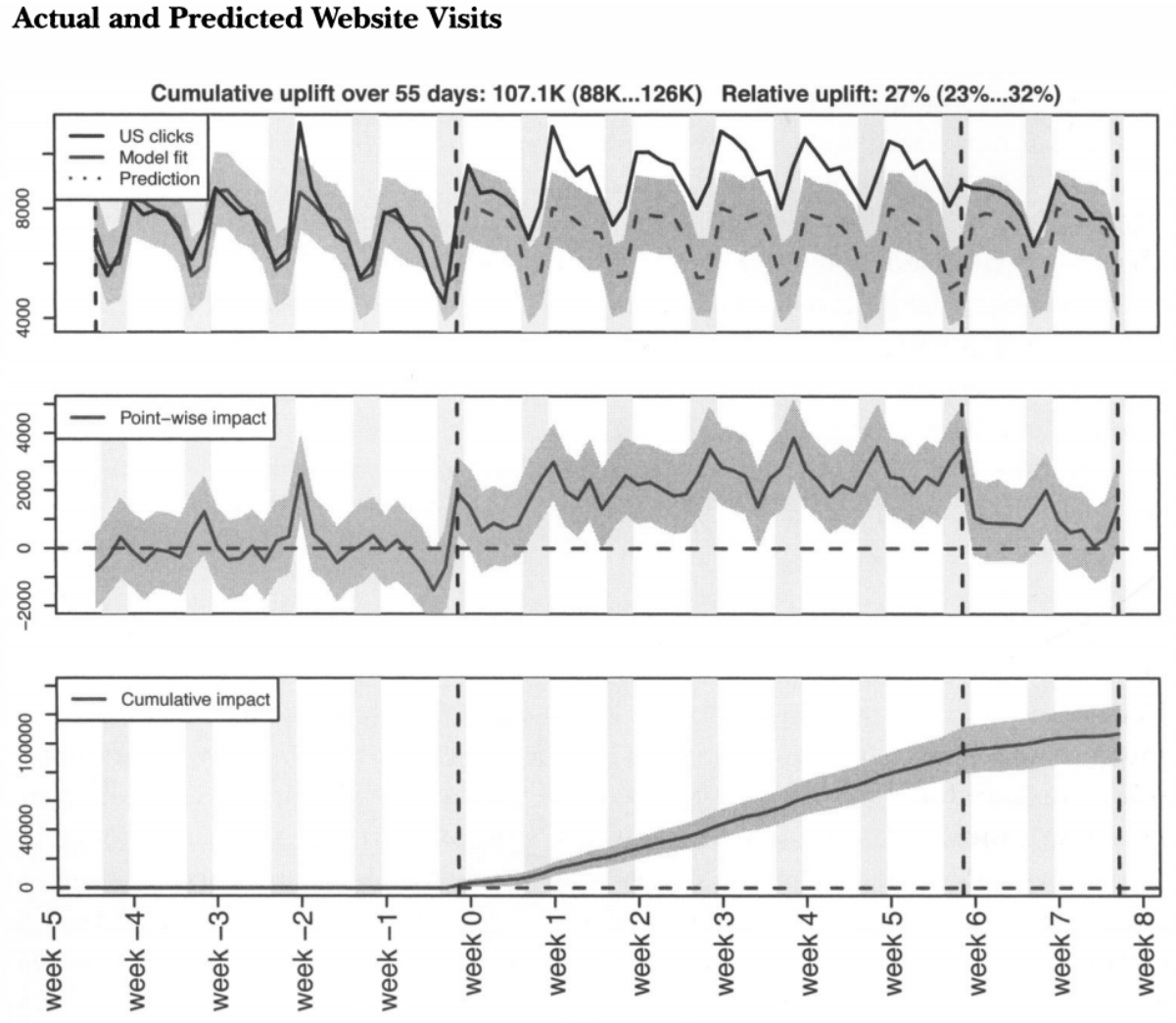


Figure: Panel A shows the actual visits and the prediction of what the visits would have been without the campaign. Panel B shows the difference, and panel c shows cumulative difference.

Model uncertainty

- Pros: averages of macroeconomic model forecasts outperformed individual models.
- Methods: blending and stacking

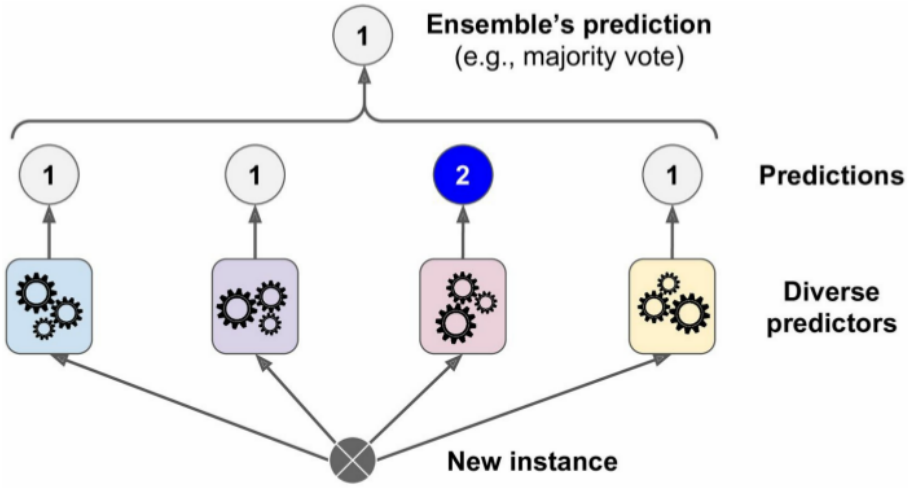


Figure 7-2. Hard voting classifier predictions

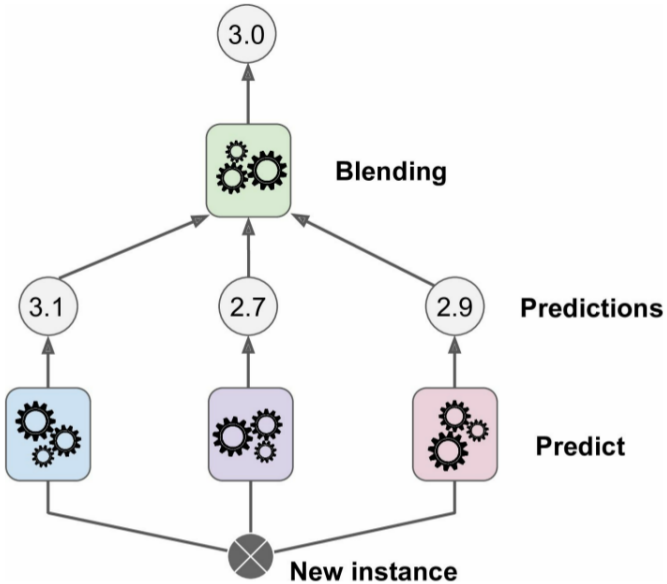


Figure 7-12. Aggregating predictions using a blending predictor

Stacking

- The first subset is used to train the predictors.
- The first layer predictors are used to make predictions on the second (held-out) set.
- Create a new training set using these predicted values as input features. Blender is trained.

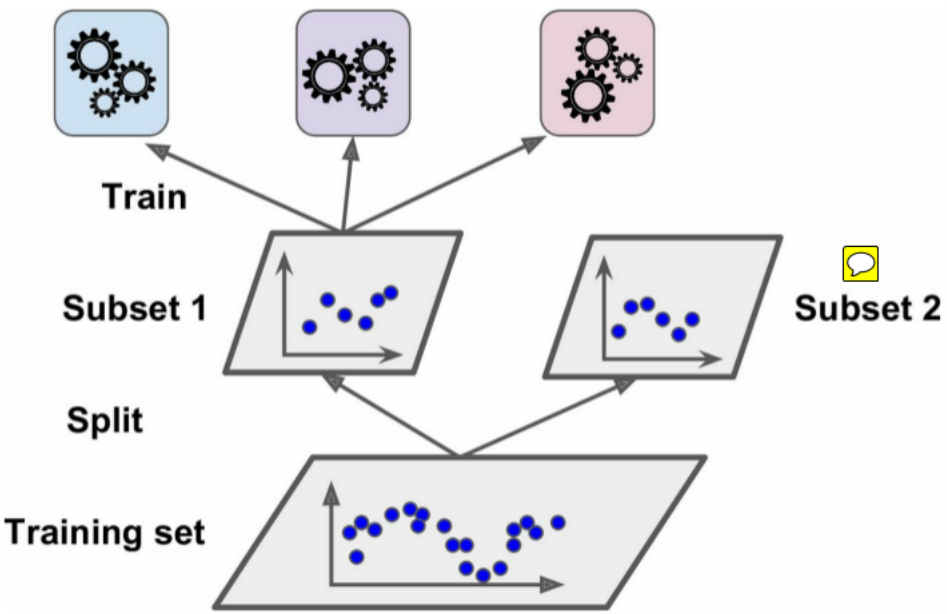


Figure 7-13. Training the first layer

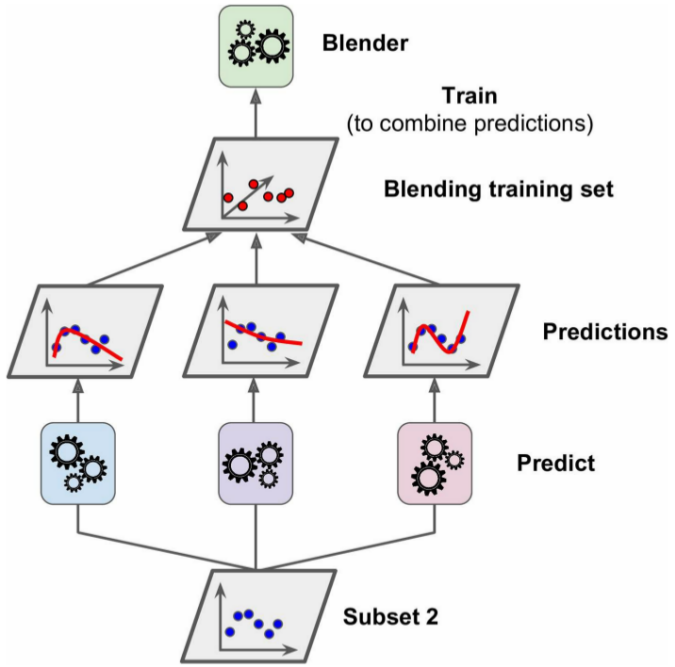
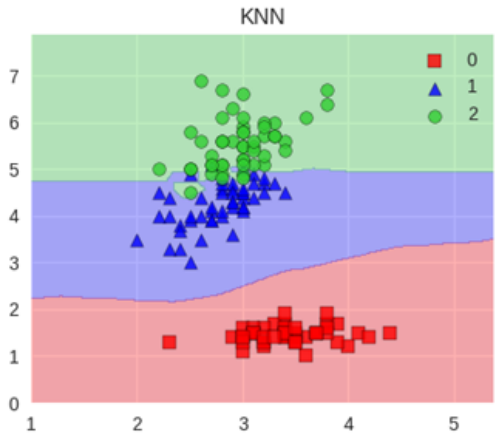
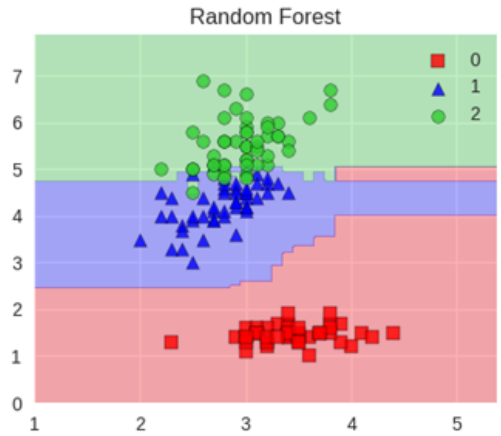
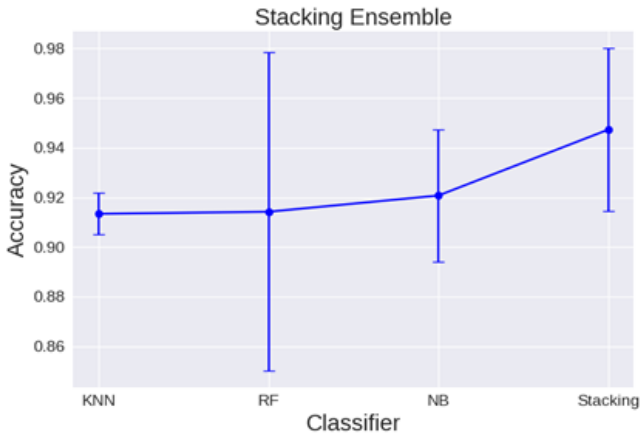


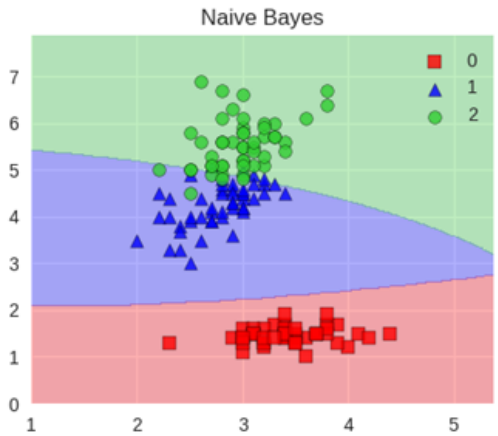
Figure 7-14. Training the blender

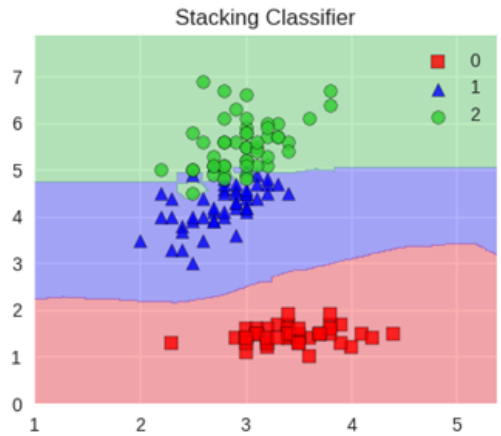
Example of stacking: (classification problem)

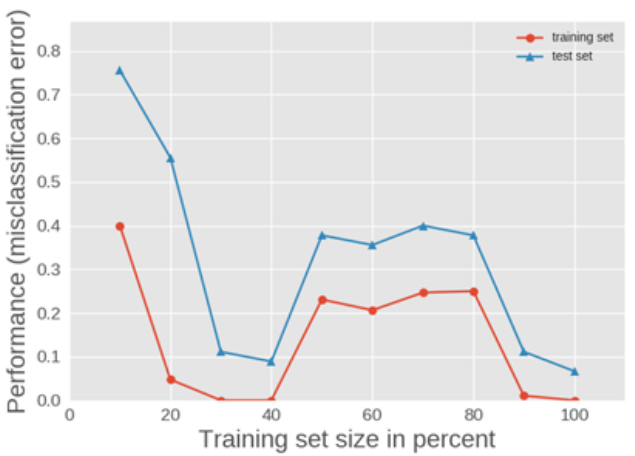


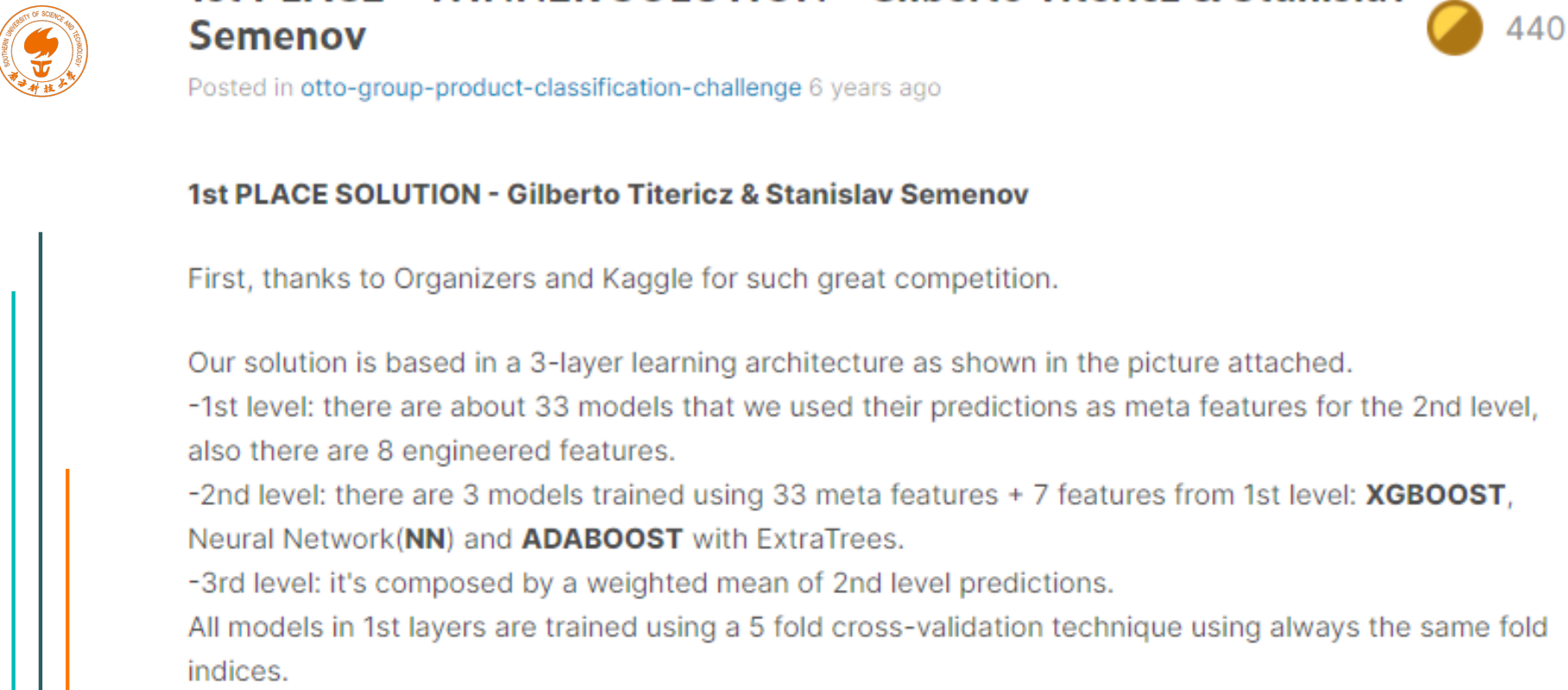












1st PLACE - WINNER SOLUTION - Gilberto Titericz & Stanislav Semenov

440

Posted in otto-group-product-classification-challenge 6 years ago

1st PLACE SOLUTION - Gilberto Titericz & Stanislav Semenov

First, thanks to Organizers and Kaggle for such great competition.

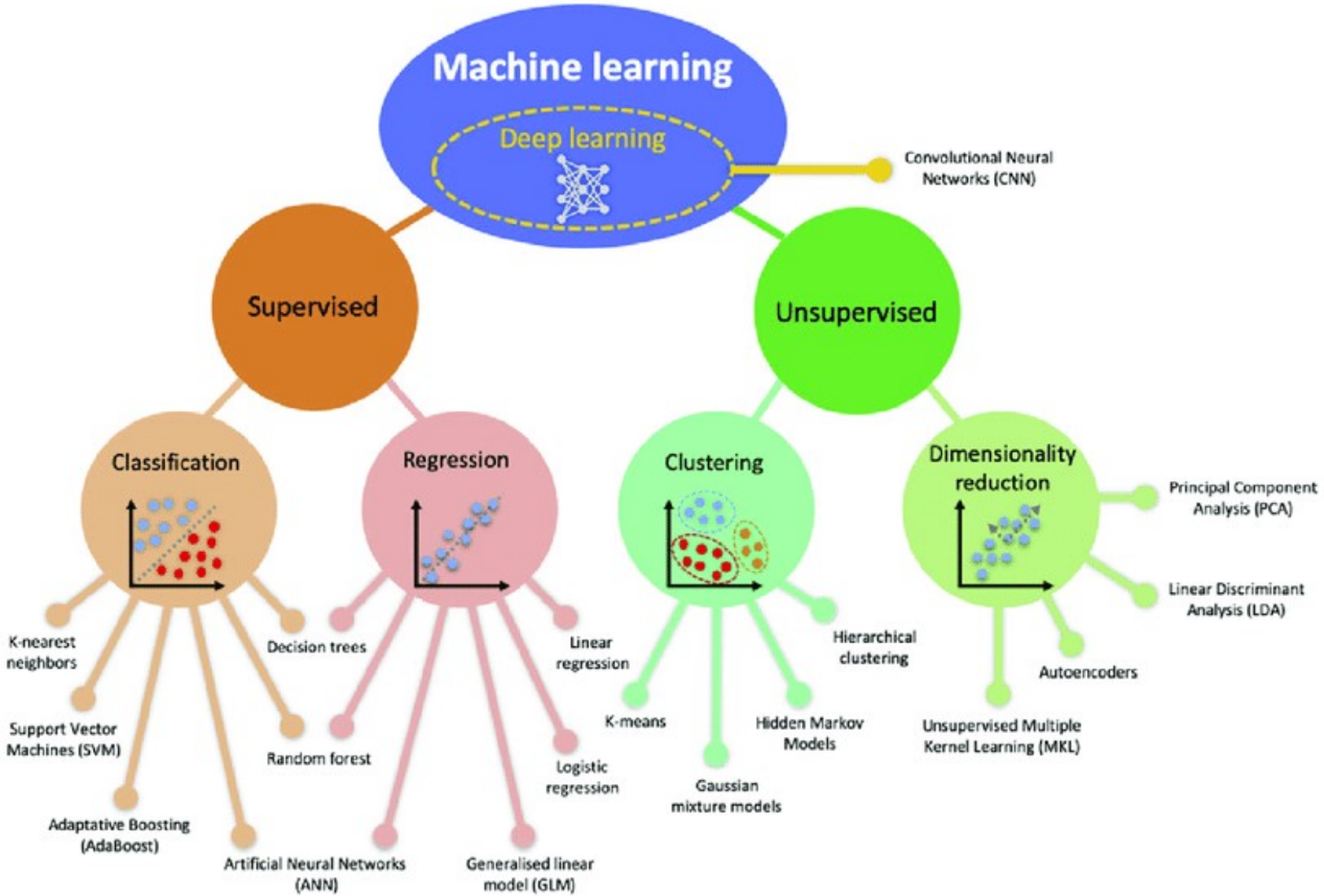
Our solution is based in a 3-layer learning architecture as shown in the picture attached.
-1st level: there are about 33 models that we used their predictions as meta features for the 2nd level, also there are 8 engineered features.
-2nd level: there are 3 models trained using 33 meta features + 7 features from 1st level: **XGBOOST**, Neural Network(**NN**) and **ADABOOST** with ExtraTrees.
-3rd level: it's composed by a weighted mean of 2nd level predictions.
All models in 1st layers are trained using a 5 fold cross-validation technique using always the same fold indices.

Figure: Stacking has been widely applied on Kaggles

Future work

- Conditions of machine learning:
  - Focus on the characteristics of data (convexity, sparsity)
  - Data preprocessing (missing data, outliers)
  - Multicollinearity between variables
- Relation to econometrics:
  - Introduce causality to some regression/classification problem
  - Balance between prediction result and interpretability
  - Combination of unsupervised learning (factor analysis)

SUSTech


Machine learning
Deep learning
Convolutional Neural Networks (CNN)
Supervised
Unsupervised
Classification
Regression
Clustering
Dimensionality reduction
K-nearest neighbors
Support Vector Machines (SVM)
Adaptative Boosting (AdaBoost)
Decision trees
Random forest
Artificial Neural Networks (ANN)
Linear regression
Logistic regression
Generalised linear model (GLM)
K-means
Gaussian mixture models
Hierarchical clustering
Hidden Markov Models
Principal Component Analysis (PCA)
Linear Discriminant Analysis (LDA)
Autoencoders
Unsupervised Multiple Kernel Learning (MKL)

References

1. Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3-28. doi:10.1257/jep.28.2.3
2. Xue Ying 2019 J. Phys.: Conf. Ser. 1168 022022
3. Mullainathan, S., Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106. doi:10.1257/jep.31.2.87

SUSTech

Thank you for listening!