Competence Estimation impact in Multi-Agent Systems

Francesca McFadden, freale1@umbc.edu
Graduate Student, expected graduation May 2026
Department of Mathematics and Statistics
University of Maryland Baltimore County

DATAWorks 2026, Washington, D.C.,
Session 5B: Advanced Modeling and Inference Techniques
Session Organizer Catherine Chalikan

4:10 – 4:40 PM
22 April 2026

Overview

- Competence scores, and similarly motivated trust scores, aim to estimate a model's suitability for prediction for a given input.
- An approach to an ensemble classification demonstrated that integrating competence estimation into ensemble learning leads to better performance compared to the highest confidence selection strategy.
- The strategy has been extended to regression base models.
- Multi agent systems incorporate the predictions and results from multiple autonomous, but coordinating agents that negotiate (or compete) to complete a task.
- MAS may employ ensemble of predictions from many specialized agents
- Agents may be using regression or classification models

The ITEA JOURNAL OF TEST AND EVALUATION

SEPTEMBER 2025 | Volume 46, Issue 3

SEPTEMBER 2025
Volume 46 | Issue 3

IN THIS JOURNAL:

Issue at a Glance
Chairman's Message
Technical Articles
Kernel Model Validation: How To Do It, And Why You Should Care
Confidence-Based Skip-Lot Sampling
Eucalyptus – An Analysis Suite for Fault Trees with Uncertainty Quantification
Digital Twins in Reliability Engineering: Innovations, Challenges and Opportunities
Competence Measure Enhanced Ensemble Learning Voting Schemes
Advancing the Test Science of LLM-enabled Systems: A Survey of Factors and Conditions that Matter Most
Beyond Accuracy: Evaluating Bayesian Neural Networks in a Real-world Application
Balancing Structure and Flexibility: Evaluating Agile, Waterfall, and Hybrid Methodologies in Aerospace and Defense Projects
Workforce of the Future
Building Confidence, Interest, and Opportunity: A Social Cognitive Career Theory-Based Analysis of the Young Women in Engineering Outreach Program
News
Association News
Chapter News
Corporate Member News

PEER REVIEWED

Competence Measure Enhanced Ensemble Learning Voting Schemes

Francesca R. McFadden
Department of Mathematics and Statistics, University of Maryland, Baltimore County; Baltimore, MD, USA

DOI: 10.61278/itea.46.3.1007

Abstract

Ensemble methods are comprised of multiple individual base models each producing a prediction. Voting schemes are used to combine, fuse, or select among the predictions of the base models. A well-formed ensemble should be formed from models with diverse assumptions, e.g., differing underlying training data, feature space selection, and therefore decision boundaries. Diversity across the models is an advantage, but is not being fully exploited with existing simple voting schemes. The purpose of the described approach is to enhance current voting scheme approaches by considering base model competence. Consideration of the individual base models in the voting for the specified input will be based on achieving a threshold model competence score. This approach appends confidence-based schemes with ensuring that inputs are consistent with the prediction space of the base models. An application using random forest classifiers demonstrates that integrating competence score estimation into ensemble learning leads to better performance compared to the highest confidence selection strategy.

Keywords: Ensemble learning, voting scheme, model competence, explainability, trust score

Introduction

Ensemble learning methods employ the predictions of multiple models to produce a single prediction. A voting scheme is applied to combine, fuse, blend, or select among the predictions of $T \geq 2$ models to produce a single prediction from the ensemble. We refer to the individual constituents of the ensemble as base models. The prediction of each base model is considered a vote. As an example of a simple majority, suppose three classifiers are applied to predict whether an image depicts a cat or a dog. The predictions 'cat' or 'dog' serve as the classifi votes – 'cat', 'cat', 'dog'. A single prediction of 'cat' will come from the ensemble.It is expec that in most cases there is consensus among the models in an ensemble, e.g., they all agree the predicted class of an input image. When there is non-consensus among the base models an ensemble, voting schemes aim to resolve it.

Classifier Model	Class A Confidence	Class B Confidence	Acceptance Threshold
Classifier 1	0.6	0.4	0.5
Classifier 2	0.3	0.7	0.8

Table 1: Example of two classifier predictions on an input data point.

To describe candidate voting schemes, we use the example shown in Table 1 where two

ISSN: 1054-0229, ISSN-L: 1054-0229
Dewey Classification: L 681 12

The ITEA JOURNAL OF TEST AND EVALUATION

About the Journal
Call for Articles
Submit your Article
Archived Issues

PERMISSIONS

ABOUT US
CHAPTERS
EDUCATION
EVENTS
MEMBERSHIP

Join us on LinkedIn to stay updated with the latest industry insights, valuable content, and professional networking!

McFadden, F. R. (2025). Competence Measure Enhanced Ensemble Learning Voting Schemes. The ITEA Journal of Test and Evaluation. https://doi.org/10.61278/itea.46.3.1007

We show an approach for enhancing multi agent systems (MAS) performance through integration of model competence estimation reporting with classification and regression model reporting.

Outline

- **Overview on Competence Estimation**
- **Review of application to Ensemble Classification**
- **An Approach for Regression Models**
  - **A Competence Score for Regression**
  - **Demonstration of Approach for Regression Models**
- **Impact of Competence Estimation integration in Multi Agent Systems**
- **Discussion**

Calibrated Trust

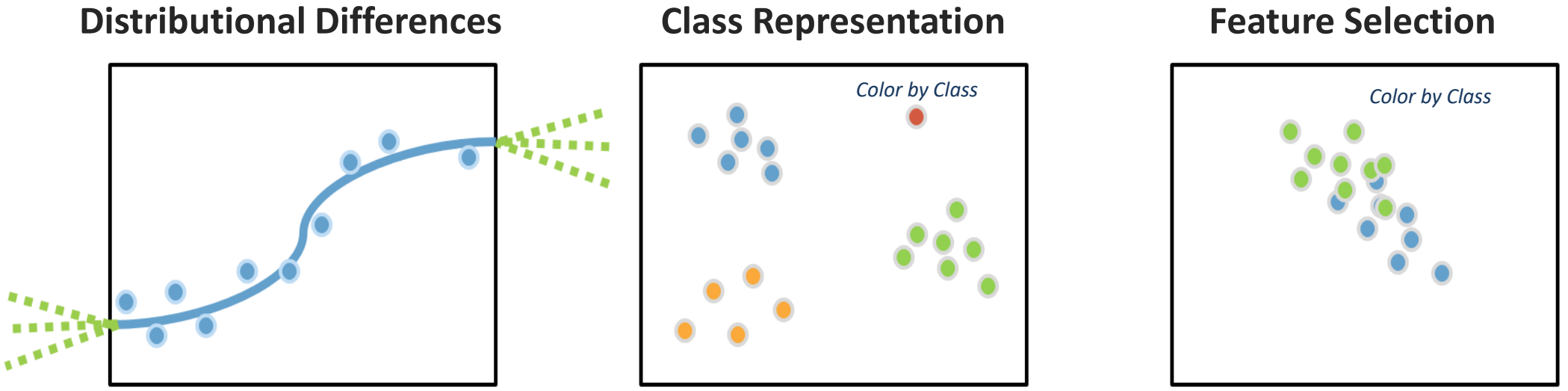
- Calibrated trust is an identified requirements theme in the MITRE Human-Machine Teaming Systems Engineering Guide.
  - Calibrated trust means that the predictions of a machine learning model are not over- or under-trusted.

- The 2019 DARPA Competency-Aware Machine Learning (CAML) program aimed to "develop machine learning systems that continuously assess their own performance in time-critical, dynamic situations and communicate that information to human team-members in an easily understood format."

How can one estimate when a machine-learned model is competent to make a prediction?

P. McDermott et al. Human-Machine Teaming Systems Engineering Guide. Tech. rep. Bedford, MA: The MITRE Corporation, 2018.
D. L. Rudd. DARPA Competency-Aware Machine Learning (CAML). 2019.

Types of Uncertainty



Depiction of types of data and model uncertainty types: (left) distributional differences (center) class representation (right) feature selection. The colors of the circles represent different true species in the data set.

- **Historically model confidence is used to estimate the effectiveness of a model's prediction.**
  - **Model confidence is incorporated in voting schemes to weigh consensus of model predictions.**
- **However, model confidence alone does not provide an indication where prediction of true class may be impacted by lack of representation in model training or possible class predictions.**
- **Uncertainty in prediction may stem from differences in the input and training data set, or model design.**
  - **Ideally, the model was trained on statistically representative data of the true population.**
  - **Differences in extent, distribution, and class representation can prevent the prediction of true class.**

ALICE (Rajendran & LeVine, 2019)

Rajendran and LeVine's approach to point-wise classification model competence estimation was introduced at the 2019 Neural Information Processing Systems (NeurIPS) conference: Accurate layerwise interpretable competence estimation (ALICE).

The ALICE score is comprised of distributional, model, and data uncertainty factors between 0 and 1.

The ALICE score for an input *x* is an indicator of whether the model will be competent to predict the true class label of an input *x*. An in-distribution factor *p(D|x)* is incorporated, where *D* is the event that *x* is in distribution. Consequently the score accounts for additional components that confidence does not.

The model is deemed competent for scores above a specified threshold. Both a correctness threshold and a risk threshold must be set based on the original definition, often requiring expert judgement.

We employed this method to estimate classification model competence in demonstrations.

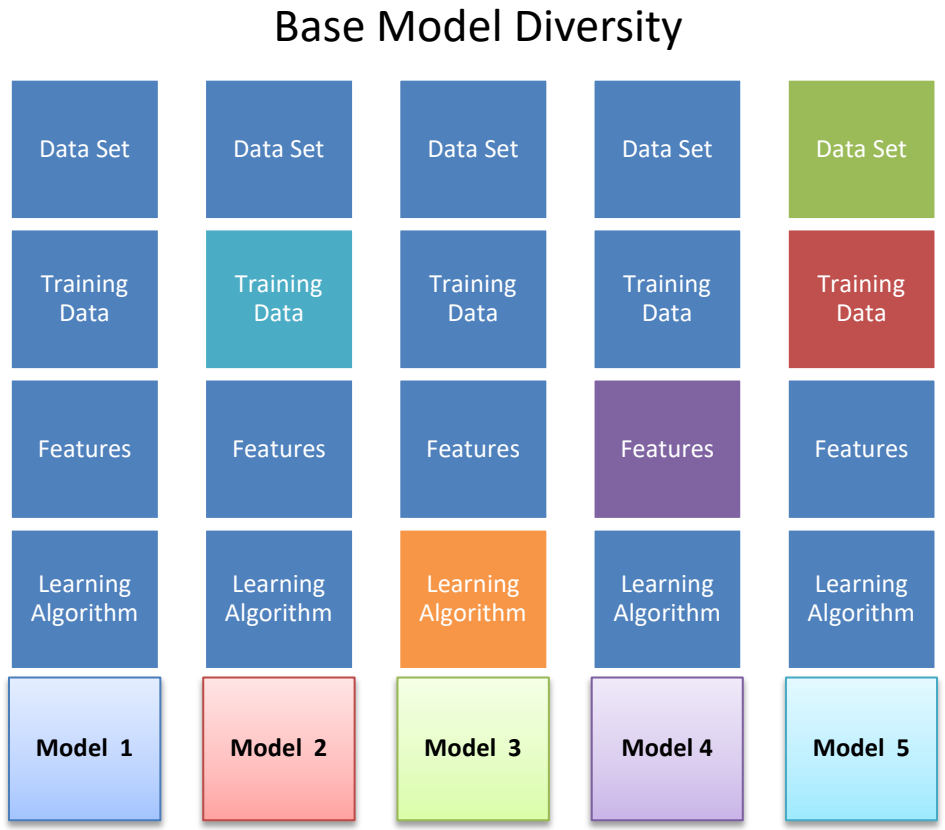
Reference: V. Rajendran & W. LeVine. Accurate layerwise interpretable competence estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, vol 32, pgs 13981–13991. Curran Associates, Inc., 2019.

Outline

- Overview on Competence Estimation
- Review of application to Ensemble Classification
- An Approach for Regression Models
  - A Competence Score for Regression
  - Demonstration of Approach for Regression Models
- Impact of Competence Estimation integration in Multi Agent Systems
- Discussion

Ensemble Learning Methods

- Ensemble learning methods combine, fuse, or select among the predictions of base models.
  - Base models may be classification models or regression models.
- A well-formed ensemble should be formed from base models with various assumptions [3], e.g.,
  - Differing underlying training data,
  - Feature space selection,
  - Learning Algorithm (NN, Decision Tree)
- As a result of these assumptions, they will therefore have differing decision boundaries.
  - Potential for differing class label space



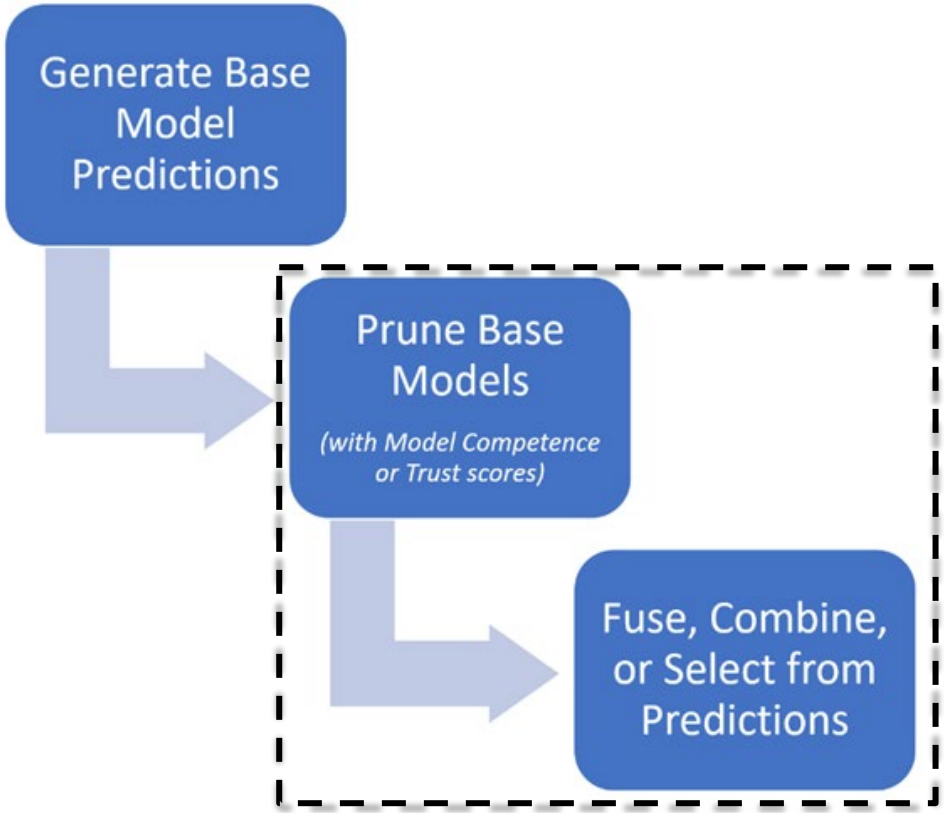
Ensemble predictions are most robust to outliers than a single base model.

Polikar, Robi, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, Third Quarter 2006.

Pruning of Base Models

- Base models may be pruned before the predictions are fused, e.g., based on having low confidence.
- In this work, we incorporate competence or trust scores into the pruning step.
- Competence or trust scores below a specified threshold will not be included when non-consensus occurs.

Only competent or trusted base learners predictions will be incorporated into the aggregate prediction.

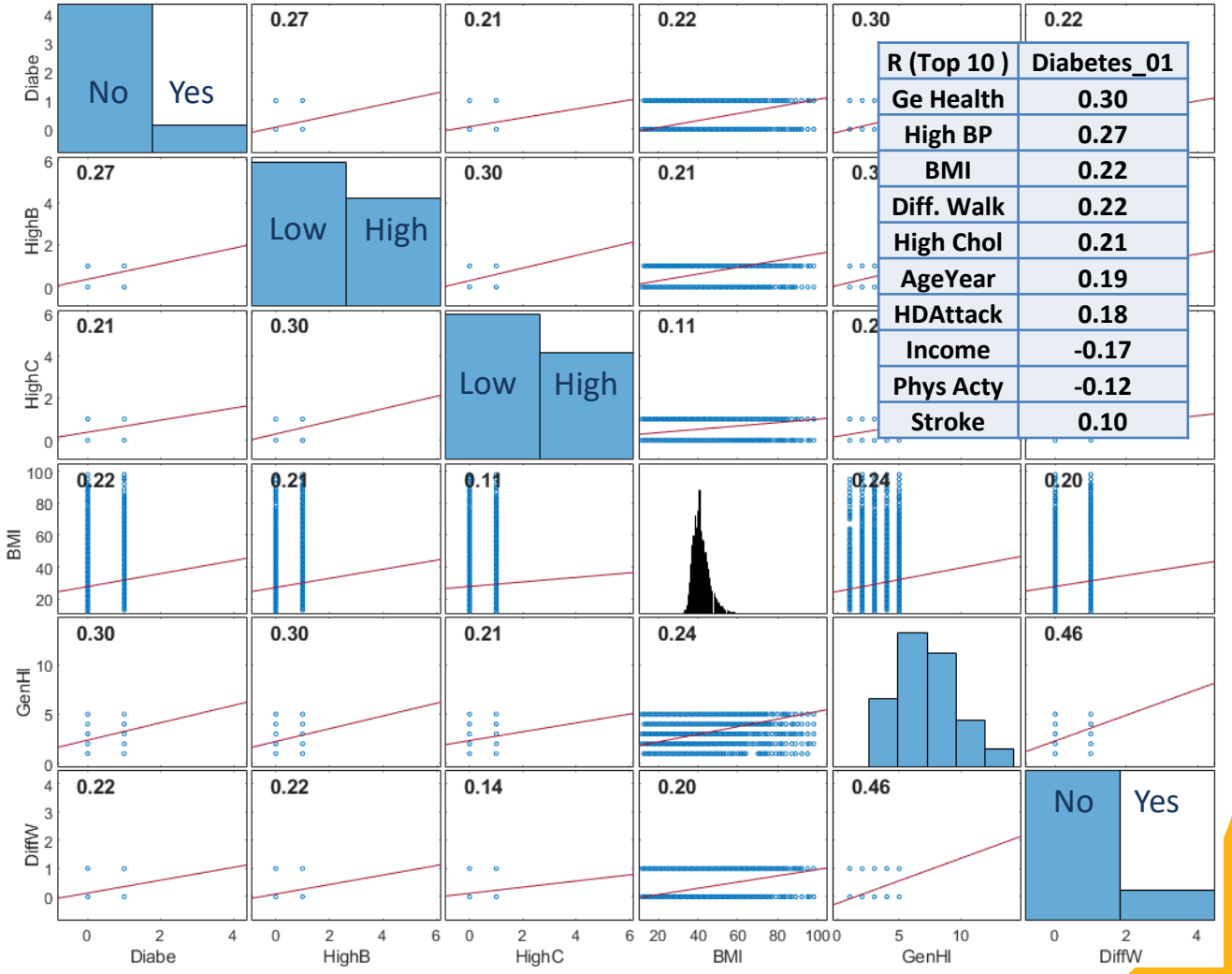


Ensemble Learning process.

Diabetes Health Indicators Dataset

- Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey [4]
  - Annual Centers for Disease Control and Prevention (CDC) survey Americans from all 50 states and 3 US territories on health-related risk factors, chronic conditions, and behaviors
  - Cleaned data set from Kaggle [5] was employed in the workflow
- 253,680 interviews with indication
  - no diabetes and/or only gestational (during pregnancy) diabetes (0)
  - prediabetes and/or diabetes (1)
- The data includes 21 features including a mixture of feature types with quantitative and qualitative responses,
  - binary, e.g., smoker or not,
  - integer, e.g., body mass index (BMI),
  - categorical scale, e.g., a general health score from 1-5; excellent to poor values

Top 5 Correlation Matrix to diabetes indicator –
high blood pressure, high cholesterol, BMI, general health, difficulty walking

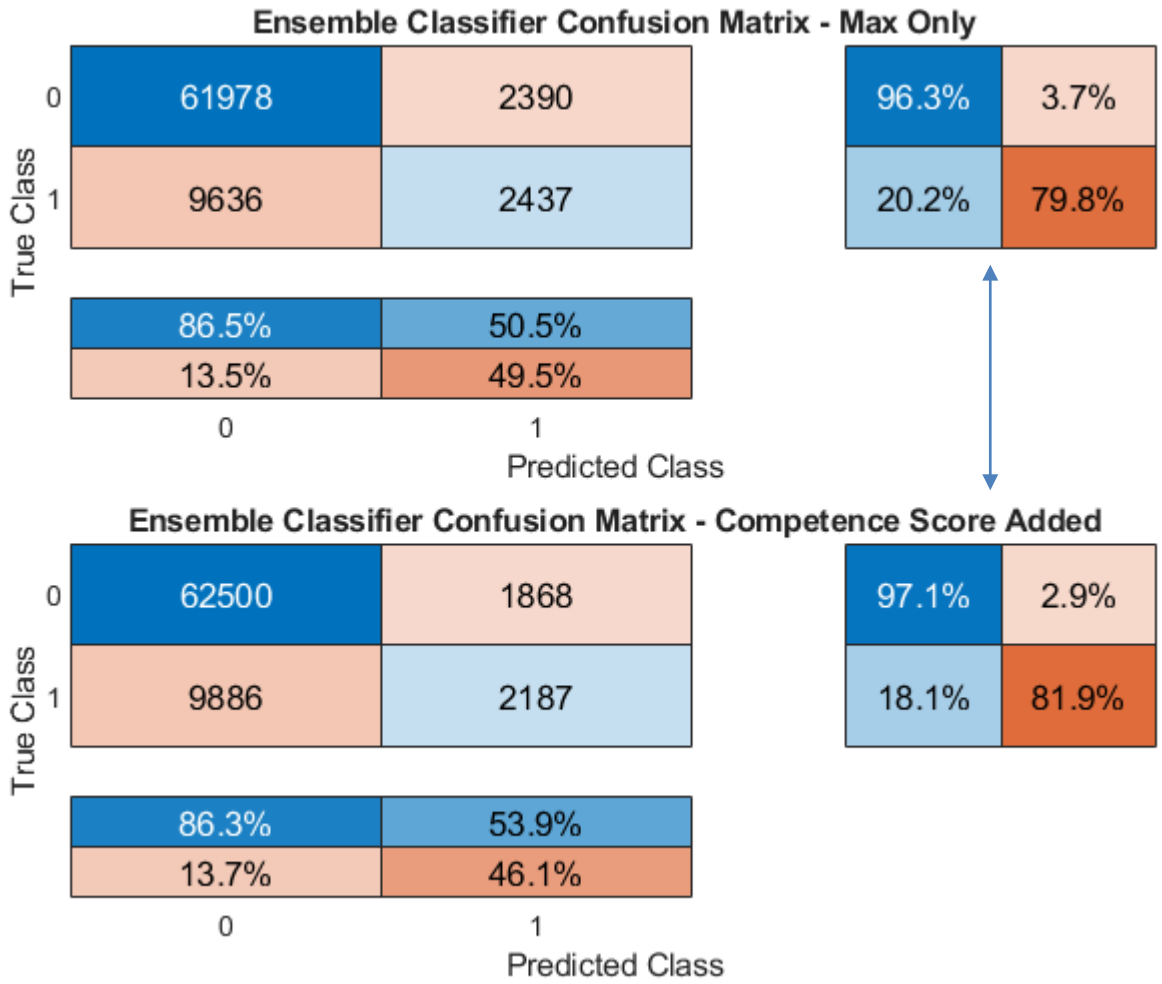


R (Top 10)	Diabetes_01
Ge Health	0.30
High BP	0.27
BMI	0.22
Diff. Walk	0.22
High Chol	0.21
AgeYear	0.19
HDAttack	0.18
Income	-0.17
Phys Acty	-0.12
Stroke	0.10

[4] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2015.
[5] Teboul, Alex. Diabetes Health Indicators Dataset, Kaggle, 2022.

Results – Ensemble Classification

- Incorporating the competence score performed slightly better than the max posterior method for true positive rate, true negative rate, false negative rate, and false positive rate.
- We were able to identify and log which classifier was used or selected for each point, leading to more transparency in selection for human machine teaming applications.



Outline

- Overview on Competence Estimation
- Review of application to Ensemble Classification
- An Approach for Regression Models
  - A Competence Score for Regression
  - Demonstration of Approach for Regression Models
- Impact of Competence Estimation integration in Multi Agent Systems
- Discussion

Ensemble Regression

- Ensemble regression combines multiple regression base models, improving robustness to outliers
- Ensemble learning methods and voting schemes suitable for classifiers, do not directly translate to regression
  - Label based approaches and corresponding confidence are not applicable
- Classic simple techniques
  - Simple averaging of base models generated via bootstrapping
  - Weighted average of base models
- More advanced
  - Training subsequent regression base models where previous ones did not perform
  - Meta models attempt to learn the diverse predictive strengths of multiple regressors

Would extending the competence enhanced voting scheme to ensemble regression also show performance value?

Considerations for a Competence Score for Regression

- Literature on competence and trust scores is more abundant for classification than it is for regression.
  - REgression TRust scOres (RETRO) [13] suggests using a weighted sum of the mean distance to neighbors and the residuals to ground truth.

- Data, distributional, model uncertainty analogous measures should be considered, but existing methods for classifiers do not necessarily translate to regressors.
  - Label based approaches and corresponding confidence are not applicable.

- An in distribution factor $p(D|x)$ remains applicable.
  - Empirical CDF of input points to individual base model training set, e.g., using Euclidean distance

- A regression model prediction interval estimates where a data point is likely to fall, accounting for the uncertainty in the model's prediction and the randomness of individual points.
  - Residuals from each base models were used to estimate 99% prediction intervals.

The combination of the latter in distribution factor and the prediction interval are applied to estimate regression competence in prediction for an individual point.

[13] de Bie, K., Lucic, A., & Haned, H. "To Trust or Not to Trust a Regressor: Estimating and Explaining Trustworthiness of Regression Predictions", arXiv.2104.06982, 2021.

Approach incorporating Competence for Regression

- Suppose $T \geq 2$ base regression models
- Consider a point x_i where $p_{ij} \approx p(D|x_i)$ for base model j
- **Model Deployment** weight base model predictions by competence

$$\hat{y}_{i,\text{InDist}} = \frac{\sum_{j=1}^{T} p_{ij}\hat{y}_{ij}}{\sum_{j=1}^{T} p_{ij}}. \tag{4}$$

- **Model Test and Evaluation** prune predictions

$$\hat{y}_{i,\text{PI}} = \frac{\sum_{j=1}^{T} \delta_{ij} p_{ij}\hat{y}_{ij}}{\sum_{j=1}^{T} \delta_{ij} p_{ij}}, \text{where} \tag{5}$$

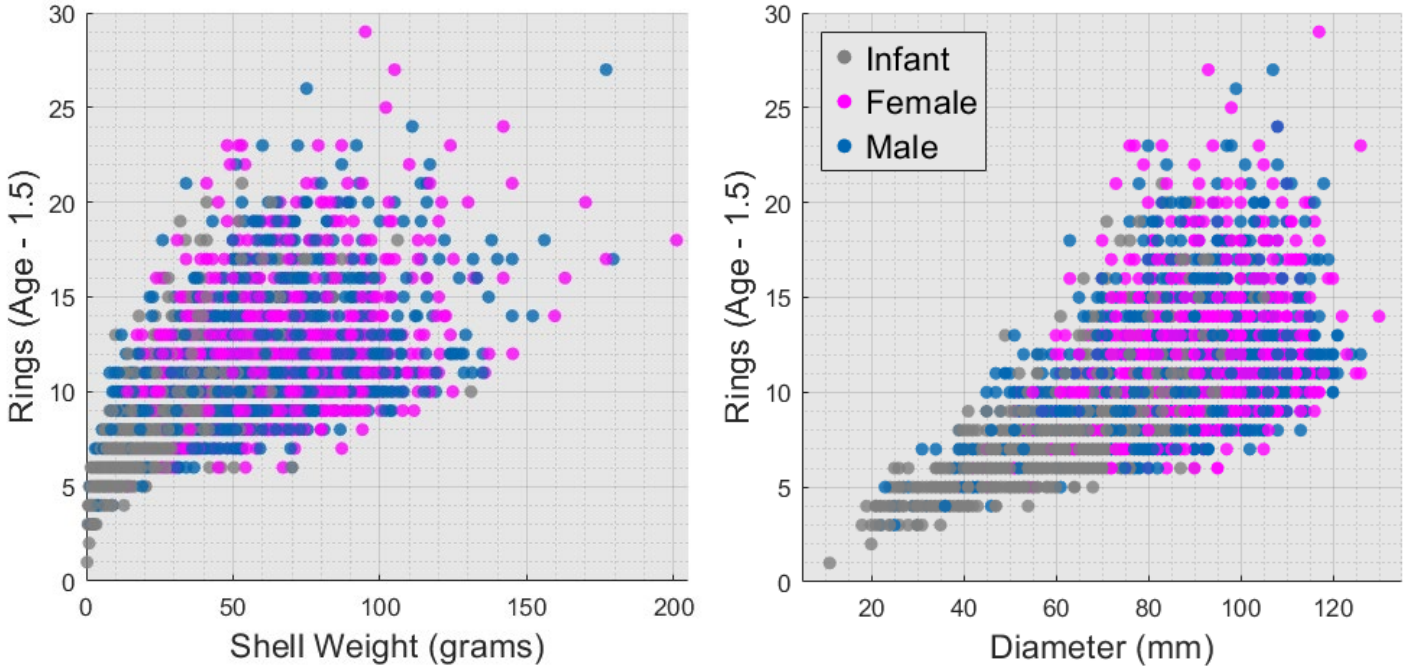
$$\delta_{ij} = \begin{cases} 1 & \text{if point } (x_i, y_i) \text{ is in the PI for base model } j \\ 0 & \text{otherwise,} \end{cases}$$

Goal: reduce mean average error (MAE) and the root mean squared error (RMSE) while preserving interpretability

Abalone data set (Nash, 1994)

Data Source: [10] Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Dataset]. UCI Machine Learning Repository.

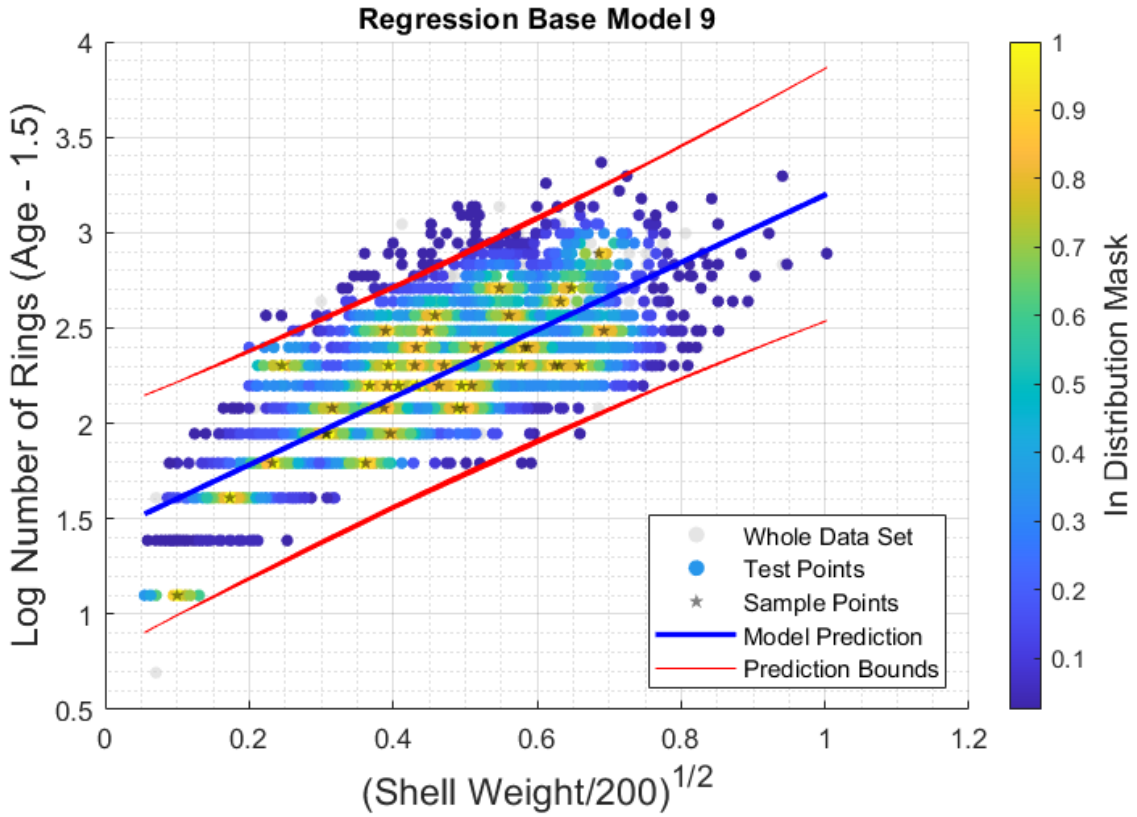
Image from caseashells.com



Measurement vs rings: Shell Weight (left) and Diameter (right)

- 4,176 Abalone snail measurements
  - Shell diameter, height, weight
  - Viscera, shucked, whole weight
  - Gender or indication of infancy
- Target: Number of Shell Rings
  - Age = # Shell Rings +1.5

Application Ensemble



- Shell weights (x_1) and the number of rings (y) were used to generate $T = 10$ regression base models
- Each model $j = 1, 2, \ldots, T$ used 40 sample points.

$$\log y_j = \beta_{0,j} + \beta_{1,j}\sqrt{\frac{x_{1,j}}{200}} + \varepsilon_j$$

Ensemble Regression Results

Table: Performance comparison of ensemble methods compared to a test set for different number of base models T and sample size N_s settings.

Ensemble Method	MAE	ΔMAE (%)	RMSE	ΔRMSE (%)
$T = 10,\ N_s = 40$				
Ensemble – Averaging	0.1724	–	0.2220	–
Ensemble – In-Distribution	0.1702	1.2406	0.2195	1.1221
Ensemble – PI Filter	0.1697	1.5135	0.2181	1.7380
$T = 10,\ N_s = 100$				
Ensemble – Averaging	0.1705	–	0.2206	–
Ensemble – In-Distribution	0.1699	0.3067	0.2199	0.3336
Ensemble – PI Filter	0.1697	0.4165	0.2194	0.5681
$T = 40,\ N_s = 40$				
Ensemble – Averaging	0.1714	–	0.2191	–
Ensemble – In-Distribution	0.1679	1.9879	0.2137	2.4822
Ensemble – PI Filter	0.1675	2.2686	0.2123	3.1313

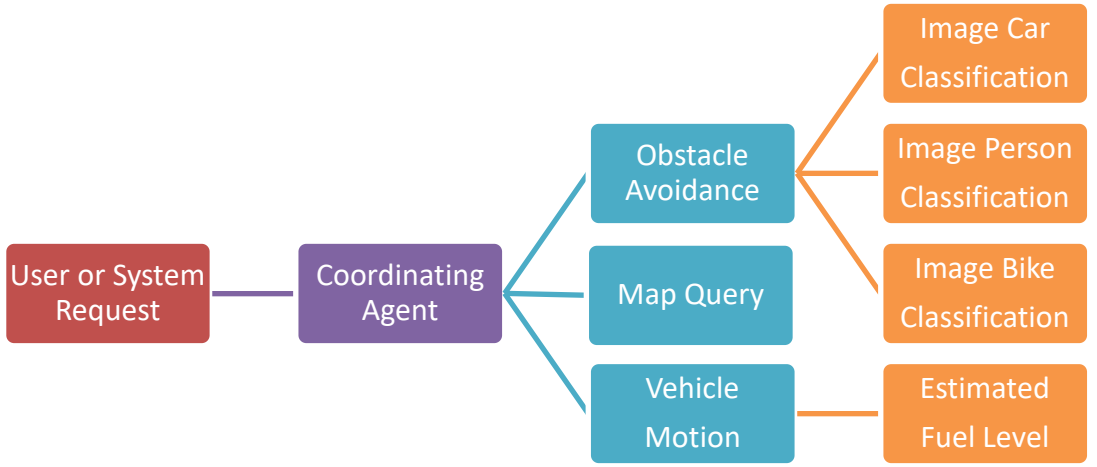
- The in-distribution weighting showed performance enhancement compared to a simple ensemble average.
- The Prediction Interval Filter further improved results.
  - Points outside of the prediction interval had lower in distribution weighting.

Outline

- Overview on Competence Estimation
- Review of application to Ensemble Classification
- An Approach for Regression Models
  - A Competence Score for Regression
  - Demonstration of Approach for Regression Models
- Impact of Competence Estimation integration in Multi Agent Systems
- Discussion

Multi Agent Systems (MAS)

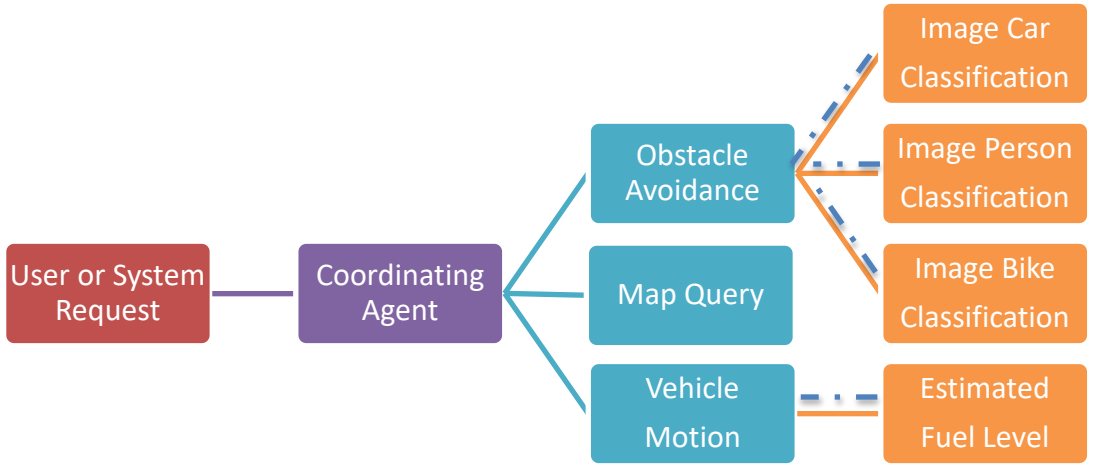
- Multi agent systems include autonomous, but coordinating agents that negotiate (or compete) to complete a task
- Agents may specialize in certain tasks, e.g., prediction in specific weather conditions
- Ensemble techniques may be used to combine agent decisions
- Agents may employ regression and classification



Example: self driving car multi agent system architecture components

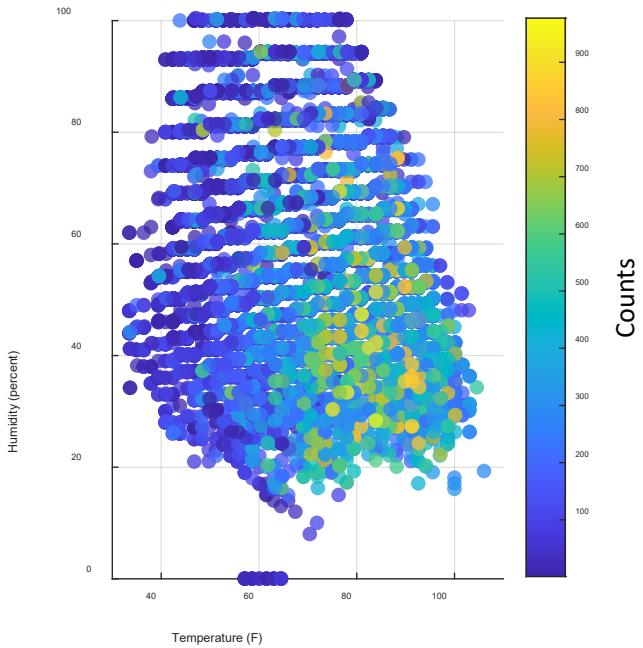
Competence integration in MAS

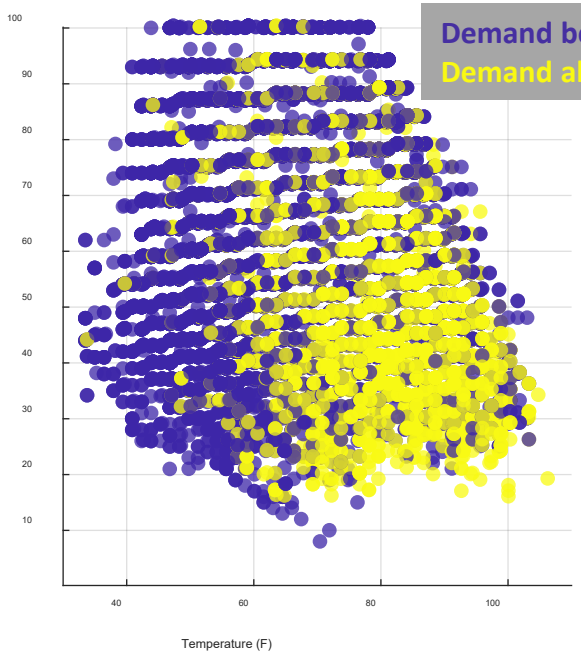
- Agents may use classification or regression models
  - Models report both prediction and competence to agents
- Agents must provide their ‘reference’ competence and prediction along with results
- Coordinating agent weighs, selects, or re-routes among agent responses using competence estimates
- Enables tailoring of expertise within a MAS to be consistent with current conditions and inputs



Example: self driving car multi agent system architecture components

Bike Sharing Data Set





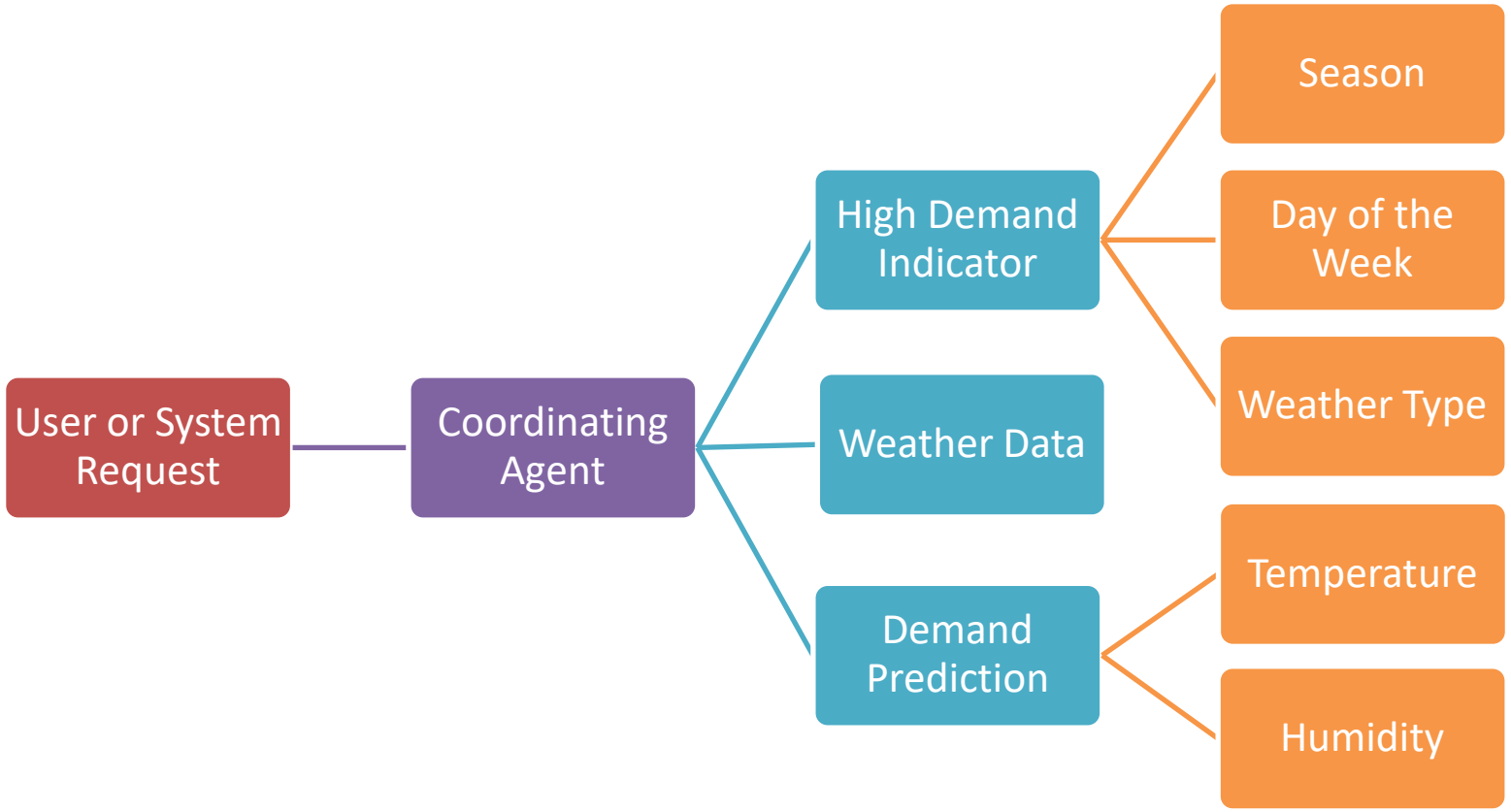
Registered Count	0.97
Casual Count	0.69
Temperature	0.40
Hour	0.39
Humidity	0.32
Year	0.25
Season	0.18
Weather Situation	0.14
Month	0.12
Wind Speed	0.09
Holiday	0.03
Work day	0.03
Week day	0.03

Absolute Pearson correlation with rental counts

H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," Progress in Artificial Intelligence, pp. 1–15, 2013.

- Bike sharing programs offer rentals which may be returned at any of multiple sites around a city, not necessarily the site the bike was acquired from.
- 17, 379 entries including variation in hour, season, weather, day, etc.
- Registered commuters and casual travelers both use these systems.
- Bike sharing rentals are highly correlated with weather data.

Application MAS



Example: bike sharing multi agent system architecture

- The bike sharing application motivates several specialized classification and regression models
- Tailor model and agent weighting based on competence estimates

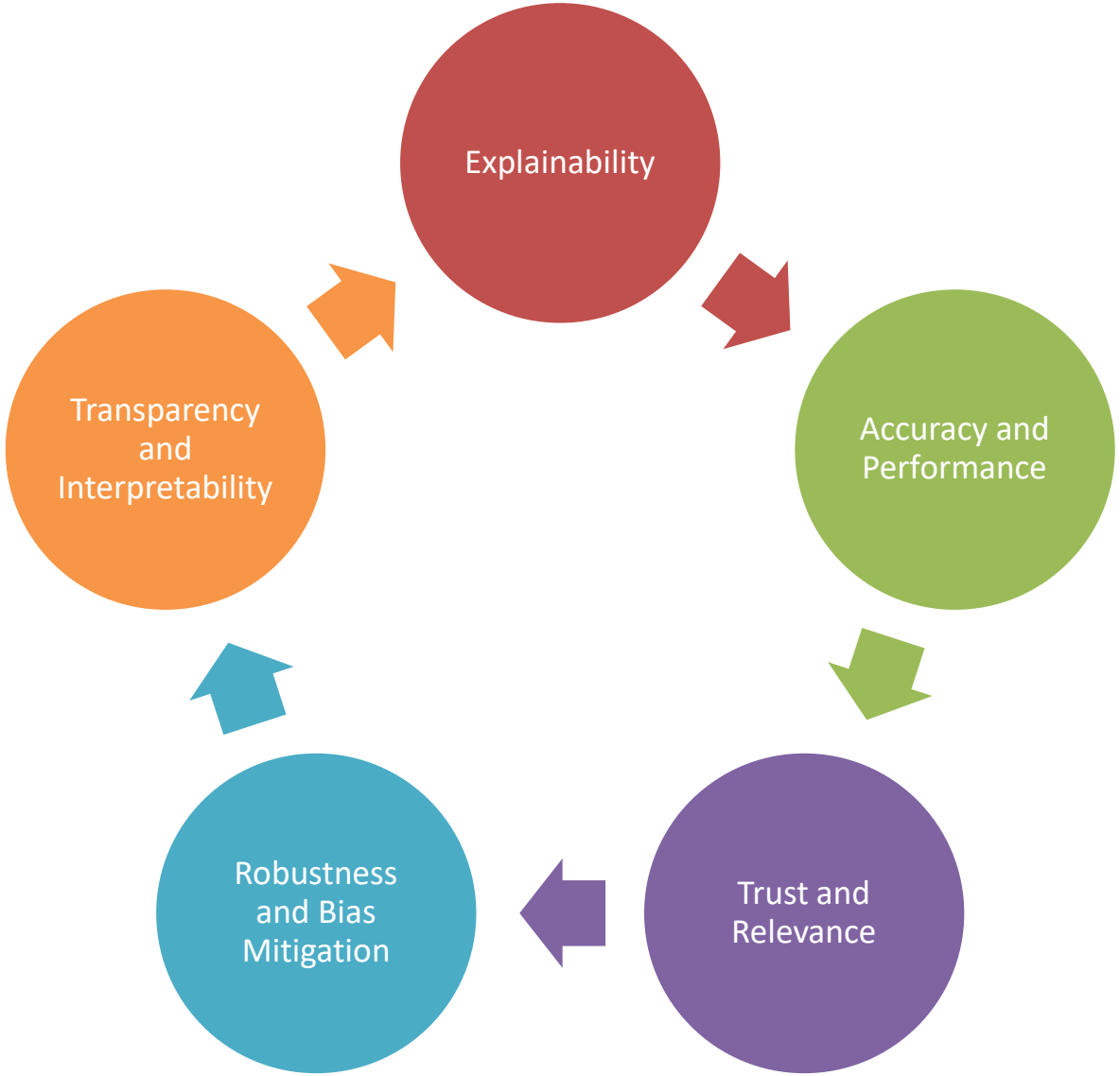
Outline

- Overview on Competence Estimation
- Review of application to Ensemble Classification
- An Approach for Regression Models
  - A Competence Score for Regression
  - Demonstration of Approach for Regression Models
- Impact of Competence Estimation integration in Multi Agent Systems
- Discussion

EXplainable Artificial Intelligence (XAI) Motivation

The combination of ensemble learning, simplistic and transparent base models, and trust or competence scores yields an XAI approach.

- Accuracy and Performance
  - Diversity among models in an ensemble will result in better prediction performance, with lower errors.
- Trust and Relevance
  - Competence and trust scores enhance recommender systems with estimation of when the model is appropriate for prediction given an input.
- Robustness and Bias Mitigation
  - An ensemble of base learners mitigates bias and hardens for robustness through its diverse composition training data and feature selection.
- Transparency and Interpretability
  - Through tracking the combination and selection of base model predictions used, the decision process is better revealed.
- Explainability
  - The approach better reveals why the ensemble produced the result.



[1] V. Rajendran & W. LeVine. Accurate layerwise interpretable competence estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, vol 32,pgs 13981–13991. Curran Associates, Inc., 2019. [2] Polikar, Robi, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, Third Quarter 2006. [3] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel). 2020 Dec 25;23(1):18. doi: 10.3390/e23010018. [4] IBM, "What is explainable AI?", https://www.ibm.com/think/topics/explainable-ai

Discussion

- Demonstrated an approach for incorporating competence score estimation into ensemble learning methods
- Described and demonstrated an extension to regressor base models [unpublished]
- Approach enables dynamic integration in both model development and deployment
  - Model competence scores may be generated at the speed of decision
- Approach is more explainable to end users than network learning ensemble techniques
  - Network approaches attempting to learn the complementary traits of base models result in loss of explainability to end users
  - From this approach recommender system visualizations may be formed to make ensemble learning with many classifiers more easily understood by end users
- Direct extension into use in multi agent systems

We extend prior work to show an approach for enhancing ensemble regression performance through integration of model competence and motivate a correspondingly approach for multi agent systems.

Compute Environment

- Processing was performed in MATLAB and Python on a 13th Gen Intel Core i7-1370P 1.90 GHz, 32.0 GB RAM, x64 processor Laptop.

- MATLAB was the environment used to process data, create classifier models, create regression models, prototype competence scores for regression models, and generate plots and confusion matrices for the results shown in this presentation

Acknowledgements

- This work was completed as an independent study at the University of Maryland Baltimore County under the advisement of Dr. Matthias K. Gobbert.

- Recognition is extended to the UC Irvine Machine Learning Repository for hosting valuable data sets that inspire prototyping and innovation in machine learning.

References

[1] V. Rajendran & W. LeVine. Accurate layerwise interpretable competence estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, vol 32,pgs 13981–13991. Curran Associates, Inc., 2019.

[2] Kundu, Rohit, The Essential Guide to Ensemble Learning, V7 Labs, 11 Jan 2024.

[3] Polikar, Robi, "Ensemble based systems in decision making," in IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[4] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2015.

[5] Teboul, Alex. Diabetes Health Indicators Dataset, Kaggle, 2022.

[6] McFadden, Francesca, “Applications of model competence estimation” [Conference Presentation], Society of Industrial and Applied Mathematics (SIAM) Mathematics of Data Science (MDS) Conference, Atlanta, GA, USA, 21-25 October 2024. https://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=80798

[7] McFadden, F. R. (2025). Competence Measure Enhanced Ensemble Learning Voting Schemes. The ITEA Journal of Test and Evaluation. https://doi.org/10.61278/itea.46.3.1007

[8] Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

[9] Moreira, João & Soares, Carlos & Jorge, Alípio & Sousa, Jorge. (2012). Ensemble Approaches for Regression: A Survey. ACM Computing Surveys. 45. 10:1-10:40. 10.1145/2379776.2379786.

[10] Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Dataset]. UCI Machine Learning Repository.

[11] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel). 2020 Dec 25;23(1):18. doi: 10.3390/e23010018.

[12] IBM, “What is explainable AI?”, https://www.ibm.com/think/topics/explainable-ai

[13] de Bie, K., Lucic, A., & Haned, H. “To Trust or Not to Trust a Regressor: Estimating and Explaining Trustworthiness of Regression Predictions”, arXiv.2104.06982, 2021.

[14] Ronald Aylmer Fisher. The use of multiple measurements in taxonomic problems. Annals Eugen., 7:179–188, 1936

Back Up