How to Get the Most Accurate Results for Your Data Using Ensembles for Numeric, Classification, Logistic and Time Series Data

Russ Conte, 2025 INFORMS Conference, Atlanta, Georgia

The vision: As easy to use as an LLM, but it’s for data.

Intro: The “Hello, world!” of ensembles

Live demo #1


2009
NETFLIX
DATE: 09-21-09
PAY TO THE ORDER OF: BellKor's Pragmatic Chaos
$1,000,000 00
AMOUNT: ONE MILLION 00/100
FOR The Netflix Prize
Reed Hastings

How much time would it take to build from scratch:

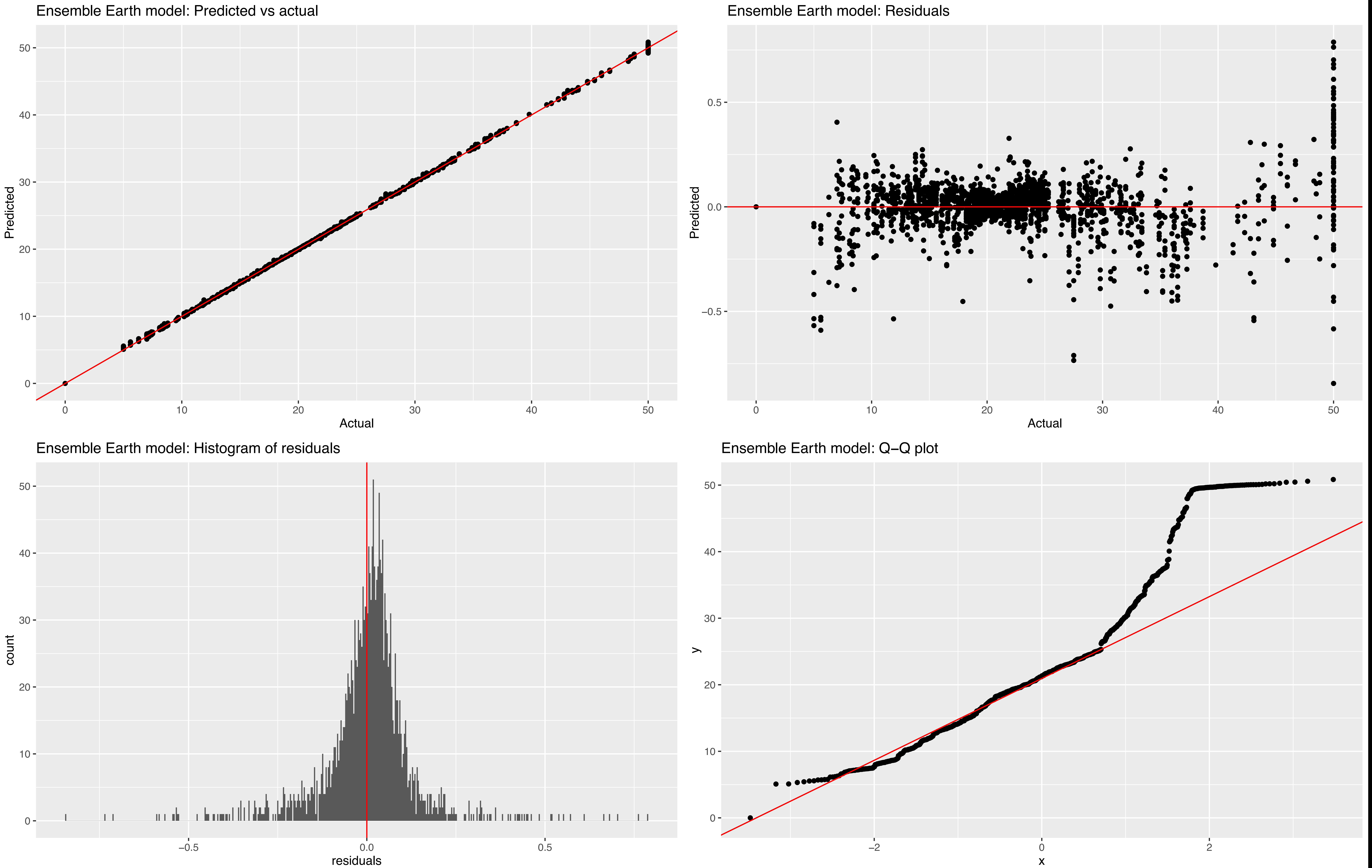
- A set of 18 individual models and 14 ensembles of models
- Automatically optimize all models which can be optimized
- Randomly resample the data as many times as the user requests
- Automatically produce all EDA, summary plots and reports
- Automatically get results that beat the best Kaggle results for 20 years
- Automatically run without any errors or warnings
- Be fully reproducible


Data
Randomize the rows
Split data into train/test/validation
Fit training data to each of the individual models
Make predictions and measure RMSE/accuracy on the holdout set
Use predictions to make an ensemble
Randomize the rows in the ensemble
Split ensemble into train/test/validation
Fit the ensemble training data to each of the ensemble models
Make predictions and measure RMSE/accuracy using the models on ensemble holdout data
Automatically create many summary reports and plots

Best results with NumericEnsembles package beats best results on Kaggle data science competitions for Boston Housing data

Kaggle Contest Name	Best Score (lowest RMSE)	# of teams	# of entries
Reitaku University	0.80946	12	32
Veronica	1.78911	41	328
UOU G03784	2.41242	49	699
SC201 June 2004	2.56893	17	402
SC201 Oct 2024	3.05221	10	220
Dupanya	3.09643	18	63
Total		147	1744
NumericEnsembles	Live demo right now	1	1

Live demo #2


Ensemble Earth model: Predicted vs actual
Predicted
Actual
Ensemble Earth model: Residuals
Predicted
Actual
Ensemble Earth model: Histogram of residuals
count
residuals
Ensemble Earth model: Q−Q plot
y
x

First (of many) pro features: How the package helps you write a professional paper using the NumericEnsembles package and helps you tell the story in your data.

Live Demo #3

Addressing the most difficult (and important) problem in data science today using the NumericEnsembles package.

Replication crisis

文A 18 languages

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **replication crisis**, also known as the **reproducibility** or **replicability crisis**, is the growing number of published scientific results that other researchers have been unable to reproduce. Because the reproducibility of empirical results is a cornerstone of the scientific method,[2] such failures undermine the credibility of theories that build on them and can call into question substantial parts of scientific knowledge.


Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

It can be proven that most claimed research findings are false.

Ioannidis (2005): "Why Most Published Research Findings Are False".[1]

The replication crisis is frequently discussed in relation to psychology and medicine, wherein considerable efforts have been undertaken to reinvestigate the results of classic studies to determine whether they are reliable, and if they turn out not to be, the reasons for the failure.[3][4] Data strongly indicate that other natural and social sciences are also affected.[5]

The phrase "replication crisis" was coined in the early 2010s as part of a growing awareness of the problem.[6] Considerations of causes and remedies have given rise to a new scientific discipline known as metascience,[7] which uses methods of empirical research to examine empirical research practice.[8]

Considerations about reproducibility can be placed into two categories. *Reproducibility* in a narrow sense refers to reexamining and validating the analysis of a given set of data. The second category, *replication*, involves repeating an existing experiment or study with new, independent data to verify the original conclusions.

Live demo #4

The vision: As easy to use as an LLM, but it’s for data.

The packages are designed to be:

Accurate
Fast
Easy to use

The goal is to make ensembles available to everyone.

No trackers, no cache, no saved data, no shared data, no LLMs, no agents, no coding assistants, no calls to any other systems.

Everything is done by the package on your local machine.

‘All truths are easy to understand, once they are discovered; the point is, to discover them.’

Galileo Galilei

Dialogue Concerning the two Chief World Systems (1632) ‘The Second Day’ tr. Stillman Drake

Solve real problems and/or make great opportunities with Ensembles for Everyone:

NumericEnsembles

ClassificationEnsembles

LogisticEnsembles

ForecastingEnsembles

All packages published on CRAN

Message me/feedback/collaborate, etc:

russconte@mac.com

Github (including slides) : InfiniteCuriosity

Russ Conte, October 28, 2025 for INFORMS Annual Meeting in Atlanta, Georgia, USA