frontiers | Frontiers in Artificial Intelligence

TYPE Hypothesis and Theory
PUBLISHED 24 February 2023
DOI 10.3389/frai.2023.1144886

OPEN ACCESS

EDITED BY
Georgios Leontidis,
University of Aberdeen, United Kingdom

REVIEWED BY
Kristina Sutiene,
Kaunas University of Technology, Lithuania
Jolita Bernatavičienė,
Vilnius University, Lithuania

*CORRESPONDENCE
Peter Taraba
✉ taraba.peter@mail.com

SPECIALTY SECTION
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 15 January 2023
ACCEPTED 06 February 2023
PUBLISHED 24 February 2023

CITATION
Taraba P (2023) Optimal blending of multiple independent prediction models.
Front. Artif. Intell. 6:1144886.
doi: 10.3389/frai.2023.1144886

Optimal blending of multiple independent prediction models

Peter Taraba*

Independent Researcher, Fort Lauderdale, FL, United States

We derive blending coefficients for the optimal blend of multiple independent prediction models with normal (Gaussian) distribution as well as the variance of the final blend. We also provide lower and upper bound estimation for the final variance and we compare these results with machine learning with counts, where only binary information (feature says yes or no only) is used for every feature and the majority of features agreeing together make the decision.

KEYWORDS

blending of independent models, normal distributions, machine learning with counts, Gaussians, going wider

Introduction

Participants of the Netflix competition used model blending heavily—refer to, for example Töscher et al. (2009), Amatriain (2013), Xiang and Yang (2009), Coscrato et al. (2020), Koren (2009), Jahrer et al. (2010), and Bothos et al. (2011). Ensemble modeling (blending) was popular not only in Netflix competition but is used also on other machine learning problems such as image processing, for example for CIFAR-10 dataset, refer to Abouelnaga et al. (2016) and Bruno et al. (2022), for the MNIST dataset refer to Ciresan et al. (2011). Ensemble modeling is used also in many other fields, for example, refer to Schuhen et al. (2012) and Ardabili et al. (2020). In this study, we derive blending coefficients based on variances of different models with only the assumption of model independence. While the formula for the final variance of the blended model and its coefficients is already derived in Kay (1993) without a proof (Equations 6.7 and 6.8 in chapter 6.4), we provide proofs both for the formula for blending coefficients and the variance of the combined model as well as the lower and upper bound estimate for the final variance based on the minimal and maximal variance of all the combined models. We also compare these results with machine learning with counts, where only binary information is used from the features to make the decision, in the last section and show very similar conclusions.

Let $\hat{y}_{k,j}$ be a prediction of model $k \in [1,N]$ for element $j \in [1,M]$, where N is the number of different independent models and M is the number of measurements we have:

$$\hat{y}_{k,j} = y_j + r_{k,j},$$

where y_j is an expected prediction and $r_{k,j}$ is a random variable with normal distribution $R_k \sim \mathcal{N}(0,\sigma_k^2)$, which has a zero average (the expected value of the variable is 0). In this study, we derive optimal blending coefficients α_k such that the blended prediction $\hat{y}_B$ is optimal:

$$\hat{y}_{B,j} = \sum_{k=1}^{N} \alpha_k \hat{y}_{k,j} = y_j \sum_{k=1}^{N} \alpha_k + \sum_{k=1}^{N} \alpha_k r_{k,j} = y_j + \sum_{k=1}^{N} \alpha_k r_{k,j}$$

with minimum variance σ_B^2, where $\sum_{k=1}^{N} \alpha_k = 1$.

Blending two independent models

Here we present two independent models

$$\hat{y}_{1,j} = y_j + r_{1,j}$$
$$\hat{y}_{2,j} = y_j + r_{2,j},$$

where $R_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $R_2 \sim \mathcal{N}(0, \sigma_2^2)$. We derive $\hat{\alpha} \in [0, 1]$ for which we get the optimal blending model

$$\hat{y}_{B,j} = \alpha(y_j + r_{1,j}) + (1-\alpha)(y_j + r_{2,j}) = y_j + \alpha r_{1,j} + (1-\alpha) r_{2,j}.$$

It is well-known fact that a random variable combining two random variables $\alpha R_1 + (1-\alpha)R_2$, where $R_1 \sim \mathcal{N}(0, \sigma_1^2)$, $R_2 \sim \mathcal{N}(0, \sigma_2^2)$ and R_1 and R_2 are independent, has a normal distribution $\mathcal{N}(0, \sigma_B^2)$, where $\sigma_B^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2$. For the mean we get:

$$\begin{aligned} E(Y_B) &= \frac{1}{M}\sum_{j=1}^{M}(\alpha r_{1,j} + (1-\alpha) r_{2,j}) \\ &= \alpha E(R_1) + (1-\alpha)E(R_2) = 0 \end{aligned}$$

and for the variance we get:

$$E(Y_B^2) = \frac{1}{M}\sum_{j=1}^{M}(\alpha r_{1,j} + (1-\alpha) r_{2,j})^2 =$$

$$= \alpha^2 \frac{1}{M}\sum_{j=1}^{M} r_{1,j}^2 + 2\alpha(1-\alpha)\frac{1}{M}\sum_{j=1}^{M} r_{1,j} r_{2,j} + (1-\alpha)^2 \frac{1}{M}\sum_{j=1}^{M} r_{2,j}^2.$$

Finally as R_1 and R_2 are independent (covariance $\frac{1}{M}\sum_{j=1} r_{1,j} r_{2,j} = 0$ is zero), we can write:

$$\sigma_B^2 = E(Y_B^2) = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2.$$

To find the optimal (we are looking for minimal value and function is convex with one minimum as we have only α^0, α^1, and α^2 dependencies—quadratic function and $\sigma_1^2 + \sigma_2^2 > 0$) blending parameter, we compute where a partial derivative of the new variance of the blended model is zero:

$$\frac{\partial \sigma_B^2}{\partial \alpha} = 2\hat{\alpha}\sigma_1^2 - 2(1-\hat{\alpha})\sigma_2^2 = 0,$$

from which

$$\hat{\alpha} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (1)$$

and the optimal variance will be:

$$\begin{aligned} \sigma_B^2(\hat{\alpha}) = \frac{\sigma_1^2\sigma_2^4}{(\sigma_1^2+\sigma_2^2)^2} + \frac{\sigma_2^2\sigma_1^4}{(\sigma_1^2+\sigma_2^2)^2} &= \frac{\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}{(\sigma_1^2+\sigma_2^2)^2} \\ &= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \end{aligned} \quad (2)$$

In Figure 1, we show how the variance is changing for different blending parameters α. Script is in Appendix 1. Blue dot—optimal value of blending parameter α matches simulation (minimal value for variance).

Blending three independent models

Now, we consider three independent models

$$\hat{y}_{1,j} = y_j + r_{1,j}$$
$$\hat{y}_{2,j} = y_j + r_{2,j},$$
$$\hat{y}_{3,j} = y_j + r_{3,j},$$

where $R_1 \sim \mathcal{N}(0, \sigma_1^2)$, $R_2 \sim \mathcal{N}(0, \sigma_2^2)$, and $R_3 \sim \mathcal{N}(0, \sigma_3^2)$.

Here, we blend optimally the first two models from the previous section:

$$\hat{y}_{4,j} = y_j + \hat{\alpha} r_{1,j} + (1-\hat{\alpha}) r_{2,j} = y_j + r_{4,j},$$

where $R_4 \sim \mathcal{N}(0, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})$ and then we find the blending parameter $\hat{\beta}$ for $\hat{y}_{3,j}$ and $\hat{y}_{4,j}$ such that

$$\hat{y}_{B,j} = y_j + \hat{\beta} r_{3,j} + (1-\hat{\beta}) r_{4,j}. \quad (3)$$

Based on the Equation (1), we get

$$\hat{\beta} = \frac{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}{\sigma_3^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2}.$$

Plugging this back into the Equation (3), we get

$$\hat{y}_{B,j} = y_j + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2} r_{3,j} + (1 - \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2}) r_{4,j}$$

$$\hat{y}_{B,j} = y_j + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2} r_{3,j} + \frac{(\sigma_1^2+\sigma_2^2)\sigma_3^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2}(\hat{\alpha} r_{1,j} + (1-\hat{\alpha}) r_{2,j})$$

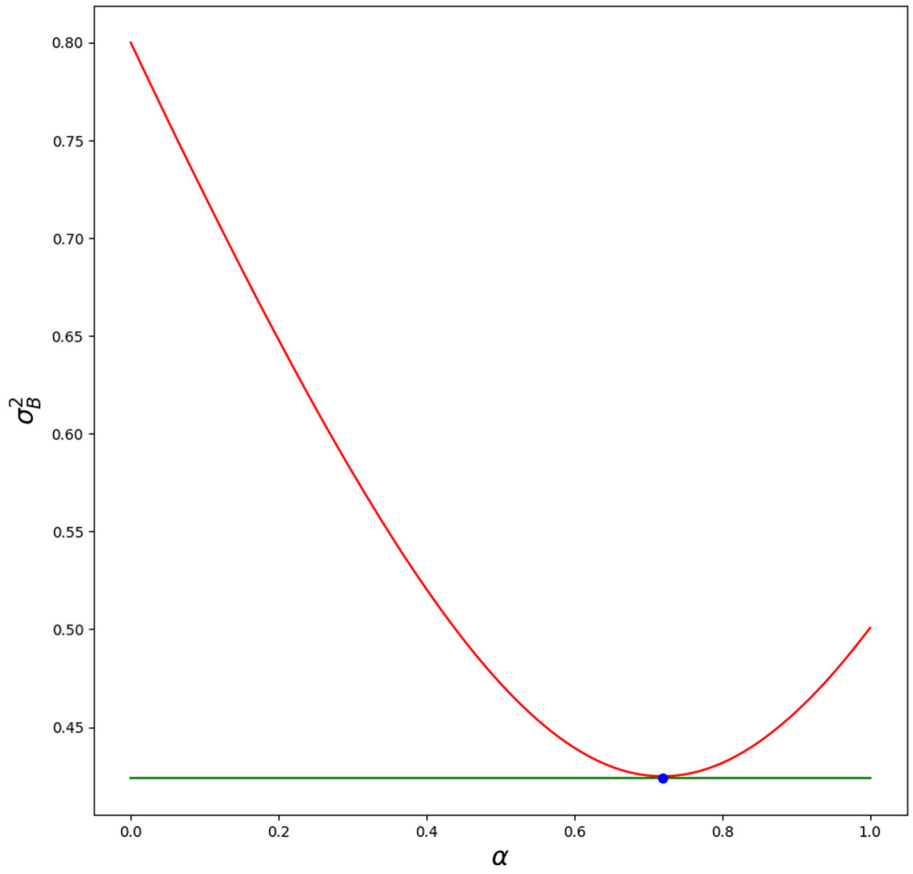

FIGURE 1
Red line—variance for different α. Green line—optimal variance σ_B^2. Blue dot—optimal α with its value $\sigma_B^2(\hat{\alpha})$. Python script is in Appendix 1.

$$\hat{y}_{B,j} = y_j + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2} r_{3,j} + \frac{(\sigma_1^2+\sigma_2^2)\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2}\left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} r_{1,j} + \left(1-\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right) r_{2,j}\right)$$

$$\hat{y}_{B,j} = y_j + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2} r_{3,j} + \frac{\sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2} r_{1,j} + \frac{\sigma_1^2\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2} r_{2,j},$$

which is symmetrical, meaning model combination order is irrelevant. Finally for $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$, we get:

$$\hat{\alpha}_1 = \frac{\sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2}$$
$$\hat{\alpha}_2 = \frac{\sigma_1^2\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2}$$
$$\hat{\alpha}_3 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2}.$$

Combining the second and third models first and then combining the result with the first model would lead to the same optimal blending parameters. The order of the combination is inconsequential. Additionally, for the final variance, we get from Equation (2)

$$\sigma_B^2(\hat{\alpha}) = \frac{\sigma_3^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}{\sigma_3^2+\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} = \frac{\sigma_1^2\sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_3^2+\sigma_2^2\sigma_3^2}.$$

In Figure 2, we show how variance is changing for different blending parameters α_1 and α_2 and $\alpha_3 = 1-\alpha_1-\alpha_2$. The script is in Appendix 2. Blue dot—the optimal value of blending parameters $(\alpha_1, \alpha_2, 1-\alpha_1-\alpha_2)$ matches the simulation (minimal value for variance).

Blending *N* independent models

Now that we have formulas for two and three different models, we prove formulas for N independent models with normal distributions:

$$\hat{y}_{k,j} = y_j + r_{k,j},$$

where $R_k \sim \mathcal{N}(0, \sigma_k^2)$. We combine these models as follows:

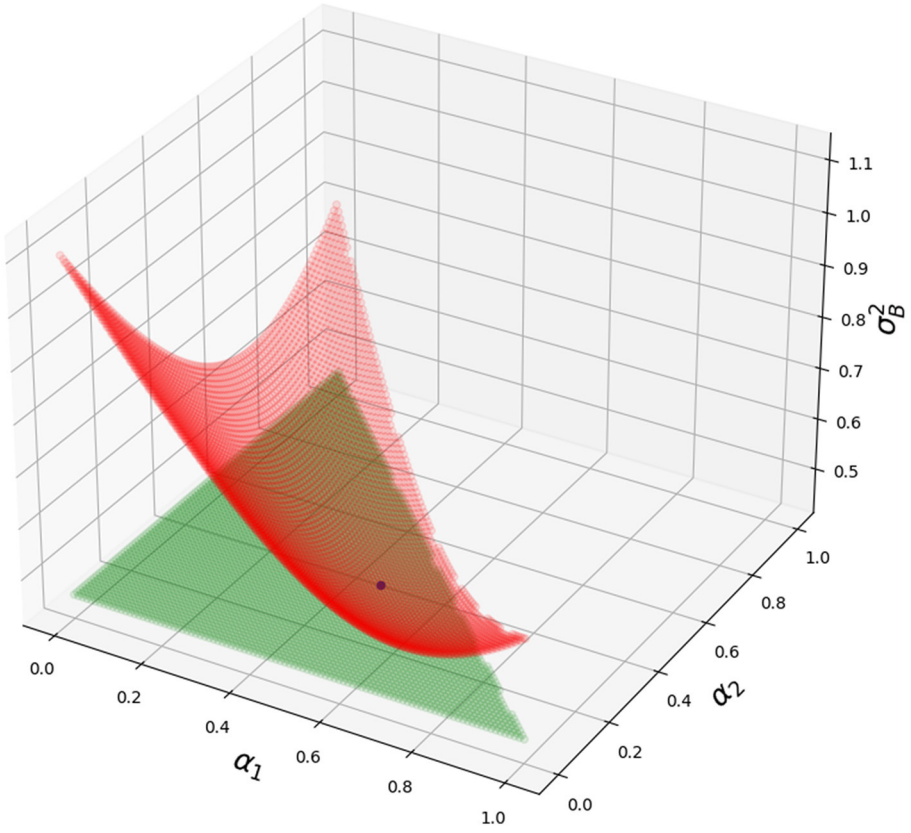

FIGURE 2
Grid consisting of red dots—variance for different α_1 and α_2. Grid consisting of green dots—optimal variance σ_B^2. Blue dot—optimal α_1 and α_2 and $1-\alpha_1-\alpha_2$ with its value $\sigma_B^2(\hat{\alpha})$. Python script is in Appendix 2.

$$\hat{y}_{B,j} = \sum_{k=1}^{N} \alpha_k \hat{y}_{k,j},$$

where $\sum_{k=1}^{N} \alpha_k = 1$ and $\alpha_k > 0$ for $k \in [1, N]$.

First, we show model independence is still needed.

Lemma 1. *Having N independent models with normal distributions $R_k \sim \mathcal{N}(0, \sigma_k^2)$ for $k \in [1, N]$, when combined as $r_{B,j} = \sum_{k=1}^{N} \alpha_k r_{k,j}$, variance of R_B is $\sigma_B^2 = \sum_{k=1}^{N} \alpha_k^2 \sigma_k^2$.*

Proof:

$$\begin{aligned}
\sigma_B^2 &= E((\sum_{k=1}^{N} \alpha_k r_{k,j})^2) \\
&= \frac{1}{M}(\sum_{k=1}^{N}\sum_{j=1}^{M} \alpha_k^2 r_{k,j}^2 + 2\sum_{k=1}^{N}\sum_{\substack{l=1 \\ l\neq k}}^{N}\sum_{j=1}^{M} \alpha_k \alpha_l r_{k,j} r_{l,j}) \\
&= \sum_{k=1}^{N} \alpha_k^2 \frac{1}{M}\sum_{j=1}^{M} r_{k,j}^2 + 2\sum_{k=1}^{N}\sum_{\substack{l=1 \\ l\neq k}}^{N} \alpha_k \alpha_l \frac{1}{M}\sum_{j=1}^{M} r_{k,j} r_{l,j}
\end{aligned}$$

As models are independent (covariance $\frac{1}{M}\sum_{j=1}^{M} r_{k,j} r_{l,j} = 0$ is zero for $l \neq k$), we get

$$\sigma_B^2 = \sum_{k=1}^{N} \alpha_k^2 \frac{1}{M}\sum_{j=1}^{M} r_{k,j}^2 = \sum_{k=1}^{N} \alpha_k^2 \sigma_k^2,$$

which ends the proof.

Theorem 2. *Having N independent models with normal distributions $R_k \sim \mathcal{N}(0, \sigma_k^2)$ for $k \in [1, N]$, we get an optimal blend with parameters*

$$\hat{\alpha_k} = \frac{\prod_{\substack{j=1 \\ j\neq k}}^{N} \sigma_j^2}{\sum_{i=1}^{N} \prod_{\substack{j=1 \\ j\neq i}}^{N} \sigma_j^2},$$

and these independent models form normal distribution $\mathcal{N}(0, \sigma_B^2)$, which has variance

$$\sigma_B^2 = \frac{\prod_{j=1}^{N} \sigma_j^2}{\sum_{i=1}^{N} \prod_{\substack{j=1 \\ j\neq i}}^{N} \sigma_j^2}.$$

Proof: For $N = 2$, we have shown it in Section 2. Now we use induction, if it is true for N, then it is true also for $N+1$.

Remark. *We have shown this also for three models in Section 3, but as for induction it is not needed, Section 3 is only a motivational section for how to derive final formulas for N models.*

We combine two normal distributions $\mathcal{N}(0, \frac{\prod_{j=1}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2})$ (assuming it is true for N) and $\mathcal{N}(0,\sigma_{N+1}^2)$. From Equation (2), (lemma 1 is incorporated in this equation) we get

$$\begin{aligned}\sigma_B^2 &= \frac{\frac{\prod_{j=1}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}\sigma_{N+1}^2}{\frac{\prod_{j=1}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}+\sigma_{N+1}^2}\\ &= \frac{\sigma_{N+1}^2\prod_{j=1}^{N}\sigma_j^2}{\prod_{j=1}^{N}\sigma_j^2+\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}\\ &= \frac{\prod_{j=1}^{N+1}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\end{aligned}$$

and hence, we have shown optimal variance is valid for $N+1$. Now we must show the same for the optimal coefficients. From Equation (1), we get

$$\hat{\alpha} = \frac{\sigma_{N+1}^2}{\frac{\prod_{j=1}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}+\sigma_{N+1}^2} = \frac{\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}$$

and hence,

$$\begin{aligned}&\hat{\alpha}_{N+1} = 1-\hat{\alpha}\\ &= 1-\frac{\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\\ &= \frac{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2-\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\\ &= \frac{\prod_{\substack{j=1\\j\neq N+1}}^{N+1}\sigma_j^2+\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2-\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\\ &= \frac{\prod_{\substack{j=1\\j\neq N+1}}^{N+1}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2},\end{aligned}$$

which proves $\hat{\alpha}_{N+1}$. Finally to show the same for $\hat{\alpha_k}$ for $k\in[1,N]$:

$$\begin{aligned}\hat{\alpha_k} &= \hat{\alpha}\frac{\prod_{\substack{j=1\\j\neq k}}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}\\ &= \frac{\sigma_{N+1}^2\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\frac{\prod_{\substack{j=1\\j\neq k}}^{N}\sigma_j^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma_j^2}\\ &= \frac{\sigma_{N+1}^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2}\prod_{\substack{j=1\\j\neq k}}^{N}\sigma_j^2\\ &= \frac{\prod_{\substack{j=1\\j\neq k}}^{N+1}\sigma_j^2}{\sum_{i=1}^{N+1}\prod_{\substack{j=1\\j\neq i}}^{N+1}\sigma_j^2},\end{aligned}$$

which ends the proof.

Going to infinity

If we can generate infinite independent models with distributions $R_i \sim \mathcal{N}(0,\sigma^2)$ (same variance), the final variance will be

$$\begin{aligned}\sigma_B^2 = \lim_{N\to+\infty}\frac{\prod_{j=1}^{N}\sigma^2}{\sum_{i=1}^{N}\prod_{\substack{j=1\\j\neq i}}^{N}\sigma^2} &= \lim_{N\to+\infty}\frac{\sigma^{2N}}{N\sigma^{2(N-1)}}\\ &= \lim_{N\to+\infty}\frac{\sigma^2}{N} = 0,\end{aligned}$$

which means we can combine all these models to get a perfect prediction with no errors. Naturally, creating an infinite amount of independent models (with covariances zero) is a difficult if not impossible task in real applications.

Theorem 3. *Having N independent models with normal distributions $R_k \sim \mathcal{N}(0,\sigma_k^2)$ for $k\in[1,N]$ and their variances $\sigma_k^2 \leq \sigma_M^2$, where σ_M^2 is their maximum variance, combining them optimally with coefficients from the theorem 2, their combined variance is $\sigma_B^2 \leq \frac{\sigma_M^2}{N}$.*

Proof: We use induction again. For $N=2$, we get

$$\sigma_B^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \leq \frac{\sigma_M^2}{2}$$

This is true as

$$\sigma_1^2\sigma_2^2+\sigma_1^2\sigma_2^2 \leq \sigma_M^2\sigma_1^2+\sigma_M^2\sigma_2^2,$$

because

$$\sigma_1^2\sigma_2^2 \leq \sigma_M^2\sigma_1^2$$

and

$$\sigma_1^2\sigma_2^2 \leq \sigma_M^2\sigma_2^2.$$

Now if it is true for N, then it is true also for $N+1$. If

$$\sigma_{B,N}^2 \leq \frac{\sigma_M^2}{N},$$

then

$$\sigma_{B,N+1}^2 \leq \frac{\sigma_M^2}{N+1}.$$

That is true as

$$\sigma_{B,N+1}^2 = \frac{\sigma_{B,N}^2\sigma_{N+1}^2}{\sigma_{B,N}^2+\sigma_{N+1}^2} \leq \frac{\sigma_M^2}{N+1},$$

because

$$N\sigma_{B,N}^2\sigma_{N+1}^2 + \sigma_{B,N}^2\sigma_{N+1}^2 \leq \sigma_M^2(\sigma_{B,N}^2+\sigma_{N+1}^2)$$

as both - this

$$N\sigma_{B,N}^2\sigma_{N+1}^2 \leq \sigma_M^2\sigma_{N+1}^2$$

and this

$$\sigma_{B,N}^2\sigma_{N+1}^2 \leq \sigma_M^2\sigma_{B,N}^2$$

are true, which ends the proof.

This proof means, that if we combine infinite independent models with distributions $R_i \sim \mathcal{N}(0,\sigma_i^2)$, where variance $\sigma_i^2 \leq \sigma_M^2$, we get variance:

$$\sigma_B^2 = \lim_{N\to+\infty} \frac{\prod_{j=1}^N \sigma_j^2}{\sum_{i=1}^N \prod_{\substack{j=1\\ j\neq i}}^N \sigma_j^2} \leq \lim_{N\to+\infty} \frac{\sigma_M^2}{N} = 0.$$

Combining infinite independent models with bounded variances from above leads to perfect prediction with variance zero.

It can be shown the same way as in Theorem 3 that combined variance is bounded also from below (as the proof is almost identical we avoid it here). If all distributions $R_k \sim \mathcal{N}(0,\sigma_k^2)$ for $k \in [1,N]$ have their variance in interval $\sigma_k^2 \in [\sigma_{min}^2, \sigma_{max}^2]$ for $k \in [1,N]$, then their combined variance will be in interval $\sigma_B^2 \in [\frac{\sigma_{min}^2}{N}, \frac{\sigma_{max}^2}{N}]$.

Similar conclusion with machine learning with counts

When it comes to using only counts (feature says yes or no only) in machine learning for predictions, as it is shown in Taraba (2021) (see section 7) on a nine-features example, we can come to the same conclusion as in the previous chapter that an infinite amount of features can lead to perfect prediction with no error. While the previous approach is statistical, machine learning with counts uses Pascal's triangle and binomial raised to infinity to show this. We use the binomial expansion

$$1 = (p+(1-p))^n = \sum_{i=0}^{n} \binom{n}{i} p^{n-i}(1-p)^i,$$

where p is the probability of features to be correct. As we want to have an odd amount of features to be able to make a decision purely on the counts (feature says yes or no), we will replace n with $2k+1$

$$1 = (p+(1-p))^{2k+1} = \sum_{i=0}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i.$$

This can be split into two parts, one with probability when the majority of features are correct $P_{correct}$ and one with probability when the majority of features are incorrect:

$$1 = (p+(1-p))^{2k+1} = P_{correct} + P_{incorrect},$$

where

$$P_{correct} = \sum_{i=0}^{k} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i$$

and

$$P_{incorrect} = \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i.$$

To show that an infinite amount of features can lead to perfect prediction, we have to show that $P_{correct}$ with the majority of features correct (at least $k+1$ of them correct) goes to 1 for all $p \in (0.5, 1]$

$$\lim_{k\to\infty} \sum_{i=0}^{k} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 1$$

We start by showing the simpler case first and that is when $p = 0.5$ then $P_{correct}$ and $P_{incorrect}$ are equal and $P_{correct} = P_{incorrect} = 0.5$. To show this, we can write

$$P_{correct,p=0.5} = \sum_{i=0}^{k} \binom{2k+1}{i} 0.5^{2k+1} = 0.5^{2k+1} \sum_{i=0}^{k} \binom{2k+1}{i}$$

and

$$P_{incorrect,p=0.5} = \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} 0.5^{2k+1} = 0.5^{2k+1} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i}$$

and those are equal as $\sum_{i=0}^{k} \binom{2k+1}{i} = \sum_{i=k+1}^{2k+1} \binom{2k+1}{i}$, because $\binom{2k+1}{i} = \binom{2k+1}{2k+1-i}$ for $i \in \{0, 1, \ldots, k\}$. As $P_{correct,p=0.5} = P_{incorrect,p=0.5}$ and their sum is 1 it follows that

$$1 = P_{correct,p=0.5} + P_{incorrect,p=0.5} = 2P_{correct,p=0.5}$$

and hence,

$$P_{correct,p=0.5} = P_{incorrect,p=0.5} = 0.5.$$

Now that we have shown what happens when $p = 0.5$, we show the main limit theorem for $p \in (0.5, 1]$.

Theorem 4. $\lim_{k\to\infty} \sum_{i=0}^{k} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 1$ *for all* $p \in (0.5, 1]$.

Proof: To show that $\lim_{k\to\infty} \sum_{i=0}^{k} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 1$ for all $p \in (0.5, 1]$, we will show instead that $\lim_{k\to\infty} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 0$ and as their sum is 1, $\lim_{k\to\infty} \sum_{i=0}^{k} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 1$ will follow.

First, we can rewrite $\sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i$ as

$$\sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i$$
$$= \sum_{i=0}^{k} \binom{2k+1}{k+1+i} p^{2k+1-(k+1+i)}(1-p)^{k+1+i}.$$

It is obvious that the numbers in the pascal triangle are decreasing when starting after the middle:

$$\frac{\binom{2k+1}{k+1+i+1}}{\binom{2k+1}{k+1+i}} = \frac{\frac{(2k+1)!}{(k+2+i)!(k-i-1)!}}{\frac{(2k+1)!}{(k+1+i)!(k-i)!}} = \frac{k-i}{k+i+2} < 1$$

for $i \in 0, 1, \ldots, k-1$, and hence, we can write

$$\sum_{i=0}^{k} \binom{2k+1}{k+1+i} p^{2k+1-(k+1+i)}(1-p)^{k+1+i}$$
$$< \binom{2k+1}{k+1} p^k (1-p)^{k+1} \sum_{i=0}^{k} (\frac{1-p}{p})^i.$$

As p is in $p \in (0.5, 1]$, then $\frac{1-p}{p} \in [0, 1)$, and hence, we can write

$$\binom{2k+1}{k+1} p^k (1-p)^{k+1} \sum_{i=0}^{k} (\frac{1-p}{p})^i$$
$$= \binom{2k+1}{k+1} p^k (1-p)^{k+1} \frac{1-(\frac{1-p}{p})^{k+1}}{1-(\frac{1-p}{p})}.$$

With that we can finally look at the original limit and write

$$\lim_{k\to\infty} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i \leq \lim_{k\to\infty} \binom{2k+1}{k+1} p^k (1-p)^{k+1} \frac{1-(\frac{1-p}{p})^{k+1}}{1-(\frac{1-p}{p})}.$$

As $\lim_{k\to\infty} (\frac{1-p}{p})^{k+1} = 0$ for $p \in (0.5, 1]$ as $\frac{1-p}{p} \in [0, 1)$, we can write

$$\lim_{k\to\infty} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i$$
$$\leq \frac{1}{1-(\frac{1-p}{p})} \lim_{k\to\infty} \binom{2k+1}{k+1} p^k (1-p)^{k+1}$$
$$= \frac{p}{2p-1} \lim_{k\to\infty} \binom{2k+1}{k+1} p^k (1-p)^{k+1}.$$

Now we look at $\lim_{k\to\infty} \binom{2k+1}{k+1} p^k (1-p)^{k+1}$. We can take a member $\binom{2(k+r)+1}{(k+r)+1} p^{(k+r)}(1-p)^{(k+r)+1}$ and compare it with $r' = r+1$ follower $\binom{2(k+r+1)+1}{(k+r+1)+1} p^{(k+r+1)}(1-p)^{(k+r+1)+1}$ by division

$$\frac{\binom{2(k+r+1)+1}{(k+r+1)+1} p^{(k+r+1)}(1-p)^{(k+r+1)+1}}{\binom{2(k+r)+1}{(k+r)+1} p^{(k+r)}(1-p)^{(k+r)+1}} = \frac{\frac{(2(k+r)+3)!}{(k+r+2)!(k+r+1)!}}{\frac{(2(k+r)+1)!}{(k+r+1)!(k+r)!}}$$
$$p(1-p) == \frac{(2(k+r)+3)(2(k+r)+2)}{(k+r+2)(k+r+1)} p(1-p)$$
$$< 4p(1-p) < 1,$$

for all $r \geq 0$ and $p \in (0.5, 1]$ as $p(1-p) \in [0, 0.25)$. This means we are multiplying a finite number $\binom{2k+1}{k+1} p^k (1-p)^{k+1}$, which is in interval $[0, 1]$, infinitely many times with a number larger or equal than 0 and smaller than 1, hence

$$\lim_{k\to\infty} \binom{2k+1}{k+1} p^k (1-p)^{k+1} = 0$$

and hence, the original limit

$$\lim_{k\to\infty} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i \leq 0$$

as well. As $\sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i \geq 0$, it follows that

$$\lim_{k\to\infty} \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^{2k+1-i}(1-p)^i = 0.$$

Remark. *It is worth mentioning that p is fixed and chosen from the interval* $(0.5, 1]$*, and we are not looking at the limit of* $p \to 0.5$*, but the limit of* $k \to \infty$*. For* $p = 0.5$*, we already know that* $P_{correct,p=0.5} = P_{incorrect,p=0.5} = 0.5$.

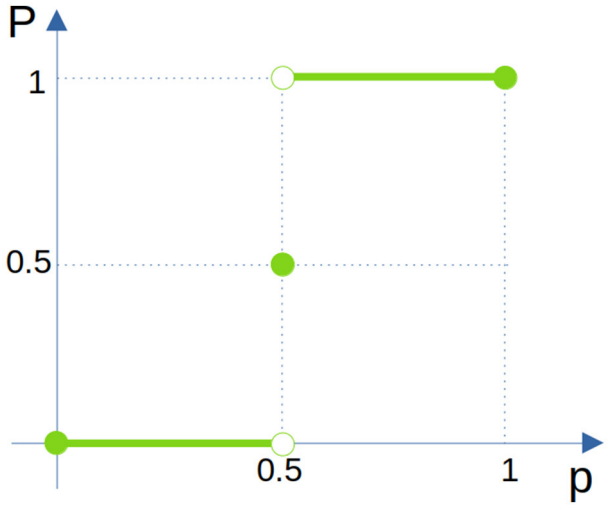

FIGURE 3
Different outcomes for $P_{correct}$ for different p intervals when k is going to ∞.

As

$$\lim_{k\to\infty}\sum_{i=0}^{k}\binom{2k+1}{i}p^{2k+1-i}(1-p)^i + \\ + \lim_{k\to\infty}\sum_{i=k+1}^{2k+1}\binom{2k+1}{i}p^{2k+1-i}(1-p)^i = 1,$$

it follows that $\lim_{k\to\infty}\sum_{i=0}^{k}\binom{2k+1}{i}p^{2k+1-i}(1-p)^i = 1$, which ends the proof.

This could be summarized for all $p \in [0,1]$ in Figure 3 as

$$P_{correct} = \begin{cases} 0 & \text{for } p \in [0, 0.5) \\ 0.5 & \text{for } p = 0.5 \\ 1 & \text{for } p \in (0.5, 1] \end{cases}.$$

This can be understood intuitively by plotting $P_{correct}$ with increasing k in Figure 4.

It is worth mentioning that with weak features with $p = 0.52$, and only $n = 15001$ of them (no need to go to infinity), $P_{correct}$ is already around 1 (see Figure 5).

As next, we show that $P_{correct}$ increases, if any of the probability of the features increases its value $p_i' > p_i$. For case $k = 0$ ($n = 1$), this is simple, as

$$P_{correct}(new) = p_1' > p_1 = P_{correct}(old).$$

For three features, it is also trivial as

$$P_{correct}(new) = p_1'p_2p_3 + p_1'p_2(1-p_3) + p_1'(1-p_2)p3 \\ +(1-p_1')p_2p_3$$

as it can be re-written to p_1' independent part and p_1' dependent part

$$P_{correct}(new) = p_2p_3 + p_1'(p_2(1-p_3) + (1-p_2)p_3)$$

and from that, it immediately follows that $P_{correct}(new) > P_{correct}(old)$ as $0 \le p_1, p_1', p_2, p_3 \le 1$ and $p_1' > p_1$. To show this for the general case, not only for cases $k \in \{0,1\}$ ($n = 2k+1 \in \{1,3\}$), we first write the general formula:

$$1 = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1}\cdots\sum_{i_{2k+1}=0}^{1}\prod_{j=1}^{2k+1}(p_j(-1)^{i_j+1} + (1-i_j)). \quad (4)$$

This can be re-written into two parts once again, $P_{correct}$, when the majority of features say yes—are correct, and $P_{incorrect}$, when the majority of features say no—are incorrect.

$$P_{correct}(p_1,\ldots,p_{2k+1}) = \sum_{i_1=0}^{1}\cdots\sum_{i_{2k+1}=0}^{1} m(i_1,\ldots,i_{2k+1}) \\ \prod_{j=1}^{2k+1}(p_j(-1)^{i_j+1} + (1-i_j)),$$

where

$$m(i_1,\ldots,i_{2k+1}) = \begin{cases} 0 & \text{for } \sum_{j=1}^{2k+1} i_j \le k \\ 1 & \text{for } \sum_{j=1}^{2k+1} i_j > k \end{cases}.$$

Theorem 5. *Increasing the probability of one of the features $p_1' > p_1$ increases final correct probability of all features, $P_{correct}(p_1', p_2, \ldots, p_{2k+1}) > P_{correct}(p_1, p_2, \ldots, p_{2k+1})$.*

Proof: We show this for increasing the first probability, as probabilities can be re-ordered and their order does not matter for the final correct probability.

Now again, as with example $k = 1, n = 3$, $P_{correct}$ can be split into p_1' independent and dependent part. The dependent part will only contain p_1' and no $(1-p_1')$ as for every $(1-p_1')$ there is also the exact same case with p_1', where even more features are correct and those two can be joined and hence it belongs to the independent part. Hence,

$$P_{correct}(p_1', p_2, \ldots, p_{2k+1}) = P_I + p_1'P_D,$$

and

$$P_{correct}(p_1, p_2, \ldots, p_{2k+1}) = P_I + p_1P_D,$$

where P_I and P_D are fixed and non-negative (as all $p_i \ge 0$ and $(1-p_i) \ge 0$) and dependent on $p_2,\ldots,p_{2k+1}$ and hence $P_{correct}(p_1', p_2, \ldots, p_{2k+1}) = P_I + p_1'P_D > P_I + p_1P_D = P_{correct}(p_1, p_2, \ldots, p_{2k+1})$ as $p_1' > p_1$, which ends the proof.

Theorem 6. *$P_{correct}(p_1, p_2, \ldots, p_{2k+1}) \ge \sum_{i=0}^{k}\binom{2k+1}{i}p_{min}^{2k+1-i}(1-p_{min})^i$, where $p_{min} = \min\{p_1, p_2, \ldots, p_{2k+1}\}$.*

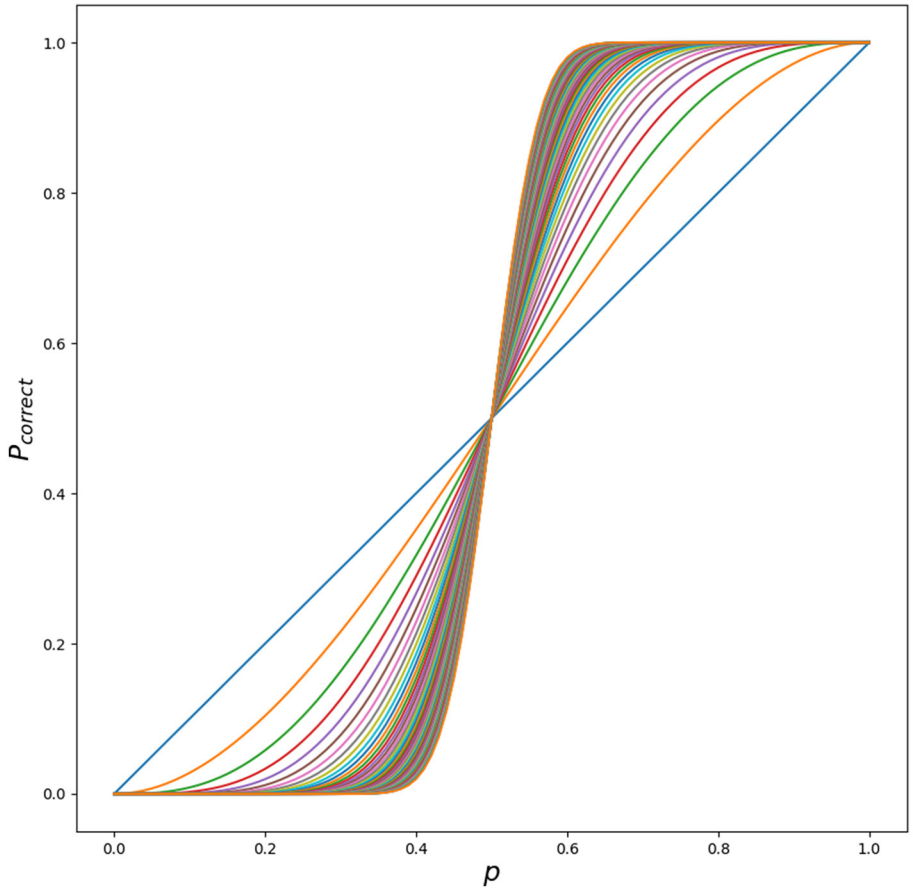

FIGURE 4
Different outcomes for $P_{correct}$ for different $p \in [0.0, 1]$ with increasing $k \in [0, \ldots, 51]$. Python script is in Appendix 3.

Proof: This proof directly follows from using the Theorem 5 multiple times by increasing every probability one at a time from the initial value p_{min}:

$$\begin{aligned} P_{correct}(p_1, p_2, \ldots, p_{2k+1}) \geq \\ \geq P_{correct}(p_{min}, p_2, \ldots, p_{2k+1}) \geq \\ \geq P_{correct}(p_{min}, p_{min}, \ldots, p_{2k+1}) \geq \\ \vdots \\ \geq P_{correct}(p_{min}, p_{min}, \ldots, p_{min}) = \\ = \sum_{i=0}^{k} \binom{2k+1}{i} p_{min}^{2k+1-i} (1 - p_{min})^i, \end{aligned}$$

which ends the proof.

Theorems 6 and 4 were needed in order to be able to say, that combining an infinite amount of features with their probabilities $p_i \in (0.5, 1]$ $(min\{p_1, \ldots, p_{2k+1}, \ldots\} > 0.5)$ will lead to perfect prediction with no error for machine learning with counts, as it was in previous Section when looking at the same problem from the statistical point of view (Gaussian distributions). It is important to say once again that these separate probabilities p_i have to be independent, as otherwise Equation (4) would not be valid.

Conclusion and discussion

We have derived blending coefficients for the ensemble of multiple independent prediction models with normal error distribution. This manuscript was mainly inspired by a Netflix competition, in which in the final stages of competition multiple teams joined their efforts to increase the accuracy of their final predictor and blending turned out to be essential to win the competition in a very short time during the final stage. This method was not only used in the Netflix competition but is used for other datasets in machine learning, such as MNIST and CIFAR-10 for image processing. We have also shown that having an infinite amount of independent predictors with their variances bounded from above is sufficient to achieve perfect prediction. While deep learning is very popular these days, one should not forget to include more features (going wider) when in need of improvement in accuracy.

Looking at a similar problem and more specifically machine learning with counts, where we only count how many features are for and against and make a decision based on a voting mechanism and a majority vote winner, we have shown once again, that an infinite amount of independent features will lead to perfect prediction when using only features which have >50% of accuracy.

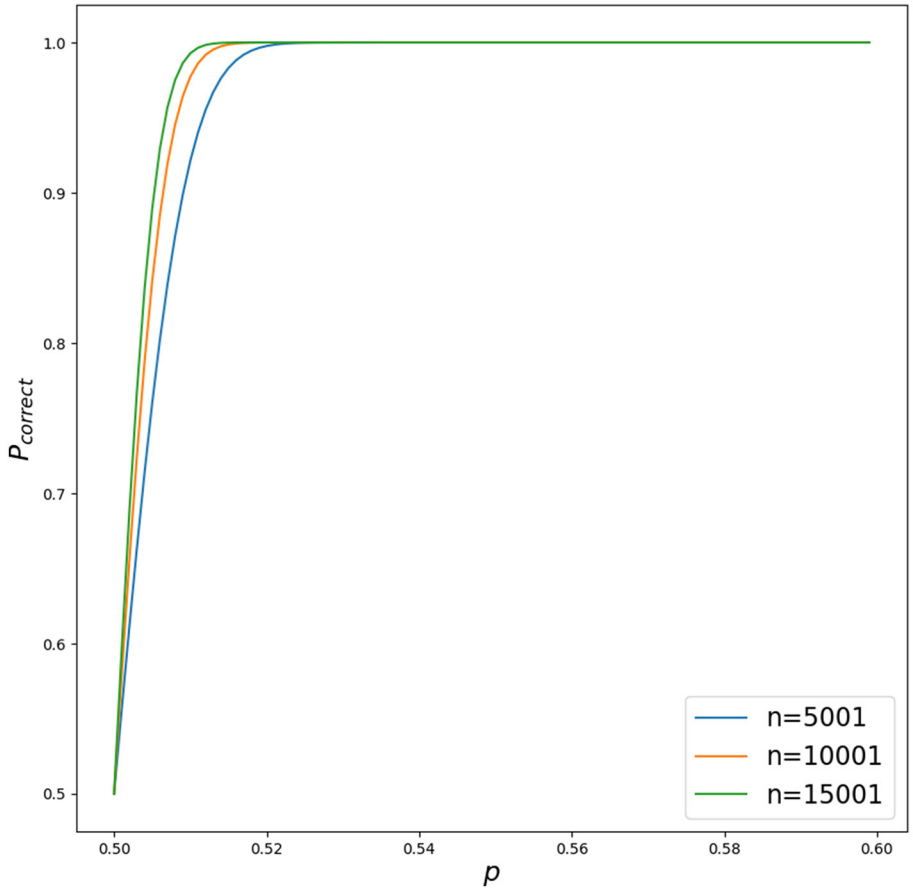

FIGURE 5
Different outcomes for $P_{correct}$ for different $p \in [0.5, 0.6]$ with increasing n for $n \in \{5001, 10001, 15001\}$. Python script is in Appendix 4.

Naturally, independent features in practice are hard to find and further the study could be made on how to convert dependent (correlated) features into independent in order to achieve as high accuracy as possible, or how to combine these dependent features together and what theoretical accuracies can be achieved. It could be also of interest how to combine features, which are not binary (for and against features), but have more than two possible outcomes.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

We would like to thank the reviewers for their suggestions and comments which led to the improvement of the manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1144886/full#supplementary-material

References

Abouelnaga, Y., Ali, O. S., Rady, H., and Moustafa, M. (2016). "Cifar-10: KNN-based ensemble of classifiers," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)* (Las Vegas, NV), 1192–1195.

Amatriain, X. (2013). "Big & personal: data and models behind netflix recommendations," in *BigMine '13* (Chicago, IL).

Ardabili, S., Mosavi, A., and Várkonyi-Kóczy, A. R. (2020). "Advances in machine learning modeling reviewing hybrid and ensemble methods," in *Engineering for Sustainable Future*, ed A. R. Várkonyi-Kóczy (Cham: Springer International Publishing), 215–227.

Bothos, E., Christidis, K., Apostolou, D., and Mentzas, G. (2011). "Information market based recommender systems fusion," in *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec '11* (New York, NY: Association for Computing Machinery), 1–8.

Bruno, A., Moroni, D., and Martinelli, M. (2022). *Efficient Adaptive Ensembling for Image Classification.* Technical report, ISTI Working Paper, 2022. Consiglio Nazionale delle Ricerche.

Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2011). "Convolutional neural network committees for handwritten character classification," in *2011 International Conference on Document Analysis and Recognition*, 1135–1139.

Coscrato, V., de Almeida Inácio, M. H., and Izbicki, R. (2020). The NN-stacking: feature weighted linear stacking through neural networks. *Neurocomputing* 399, 141–152.

Jahrer, M., Töscher, A., and Legenstein, R. (2010). "Combining predictions for accurate recommender systems," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10* (New York, NY: Association for Computing Machinery), 693–702.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc.

Koren, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix Prize Docu.* 81, 1–10.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Month. Weath. Rev.* 140, 3204–3219. doi: 10.1175/MWR-D-12-00028.1

Taraba, P. (2021). Linear regression on a set of selected templates from a pool of randomly generated templates. *Mach. Learn. Appl.* 6:100126. doi: 10.1016/j.mlwa.2021.100126

Töscher, A., Jahrer, M., and Bell, R. M. (2009). The bigchaos solution to the netflix grand prize. *Netflix Prize Docu.* 1–52.

Xiang, L., and Yang, Q. (2009). "Time-dependent models in collaborative filtering based recommender system," in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1* (Milan), 450–457.