# Side information as an organizing principle for understanding model quality

Benjamin W. Domingue[1,†], Klint Kanopka[1], and Charles Rahal[2]

[1]Graduate School of Education, Stanford University
[2]University of Oxford
[†]Correspondence about the paper should be sent to ben.domingue@gmail.com.

**Abstract**

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer quis turpis id magna aliquam rhoncus in quis purus. Maecenas tristique ut erat eget ultrices. Nunc eu sollicitudin risus. Praesent sollicitudin, justo auctor fringilla cursus, odio tellus consequat lorem, vitae fringilla ipsum lorem non justo. Fusce ultrices iaculis mauris. Sed accumsan efficitur semper. Sed varius risus venenatis lacus imperdiet, nec imperdiet est gravida. Aliquam sed ullamcorper lorem.

## 1 The concept of the IMV

**the gloomy prospect?**

The Intermodel Vigorish [1] is an approach to understanding model quality based on an equating of the different between two predictive approaches to the expected profits associated with a gamble in which one party has side information not available to the other party. While it was originally designed as a means for understanding the fit of predictive models for binary outcomes, we argue that it can also be used across a broader range of outcomes. There are several technical issues that will need to be resolved to do this, but, can it be shown to work, there are potentially large upsides as it will allow for more precise comparison of prediction quality across outcome type. Further, we think that the intuition of the metric (i.e., as the expected profit from a wager) emphasizes focus on the quality of our predictions rather than more traditional focus on whether, say, a given covariate is significant.

Construction of the IMV for a given outcome is a three-step process:

1. Translation of predictions to statements about random variables via entropy.
2. Construction of a fair bet based on the baseline model.
3. Calculation of expected profits based on side information.

We now further discuss several ideas relevant to this process.

The notion of entropy stems from Shannon's work in information theory [2]. Entropy is a measure of uncertainty in a system. Here, we use entropy to map a system of predictions to a random variable with equivalent levels of uncertainty; highly predictive models will thus be mapped to random variables with relatively small amounts of uncertainty. Entropy for a binary random variable $X$ with observations $x_i$ is defined as

$$H(X) = -\sum_i P(x_i) \log P(x_i). \tag{1}$$

We also utilize the related definition of differential entropy, $H(X) = -\int_X \log P(x)$, for use with continuous $X$.

The IMV is always a comparison of two models; in particular, we are looking at the expected profit associated with prediction via the *enhanced* model relative to the *baseline* model. The emphasis on change here is value as it focuses attention on the thing we often care most about (i.e., absolute statements about model fit frequently require further manipulation so as they can be interpreted as change in model fit). Note two things. First, our choice of terminology is meant to help intuition but it need not in fact be the case that the enhanced model improves on the baseline model. Second, it might seem restrictive to require two models. We assert that it is not given that a simple model (e.g., the prevalence of the outcome in the case of a binary outcome) can always be inserted as the baseline model.

When considering binary outcomes, there was no ambiguity about the nature of the bet given the simplicity of Bernoulli random variables. The key challenge when considering a broader array of outcome types will be construction of a bet. [why we think it still worthwile]

# 2 The IMV with continuous outcomes

Suppose we have a model $F$ for some outcome $y$ based on predictors $\mathbf{x}$ such that we are assuming

$$y_i \sim (F(\mathbf{x_i}), \sigma_i^2). \tag{2}$$

The entropy of a normal distribution, denoted $\mathrm{N}(\mu, \sigma^2)$, depends only on the variance: $\frac{1}{2}(1 + \log 2\pi\sigma^2))$. We'll denote the density for this random variable as $f_\sigma$ where we use the subscript to emphasize the variance for the normal distribution under consideration. We thus solve

$$\sum_i f_\sigma(F(x_i) - y_i) = \frac{1}{2}(1 + \log 2\pi\sigma^2)) \tag{3}$$

for $\sigma$. The intuition here is fairly straightforward: if $|F(x_i) - y_i|$ is generally small, this will lead to an induced normal distribution with small variance.

The challenge here pertains to a specification of the bet. The simplicity of the binary case left no uncertainty as to the nature of the bet. Here, things are more subtle. We consider several possibilities and then also demonstrate how this framework can be used to facilitate more straightforward comparisons of models across outcome types.

## 2.1 Bets based on $\delta$

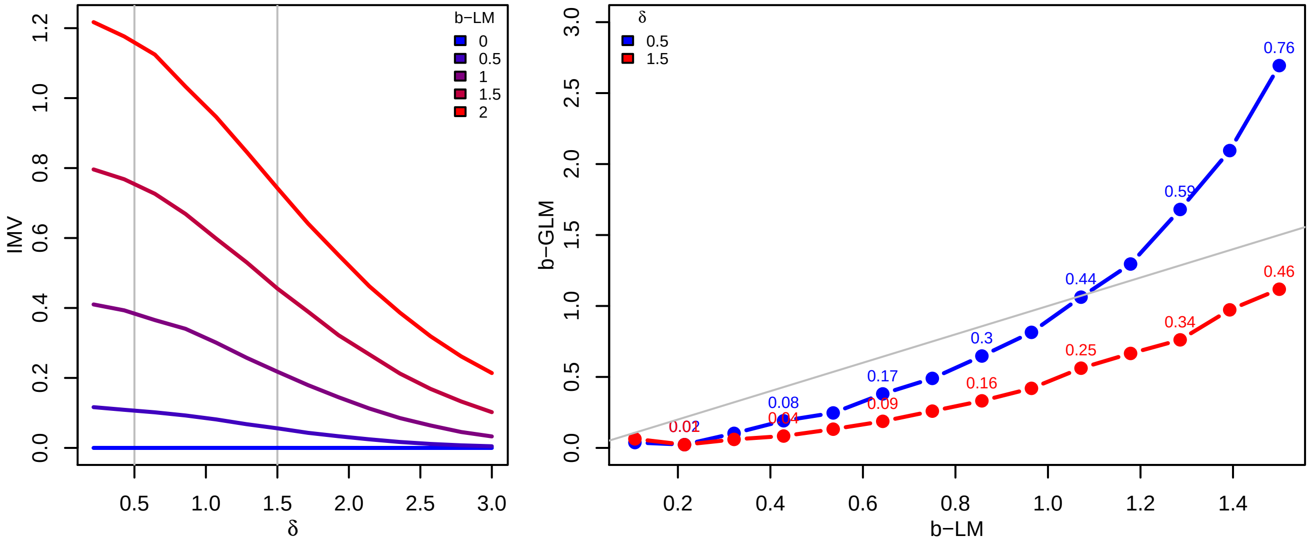
Supposing that $N(0, \sigma_0^2)$ is the normal distribution induced by the baseline prediction, we begin by making a bet based on whether a draw from the this distribution being within $\delta$ of 0. Of course, a range of choices for $\delta$ is possible; we comment on this fact momentarily. Suppose that we place a \$1 bet on a draw from this distribution being within $\delta$ of the origin. What would then be an even or fair bet for the other party? If $p_0 = \int_{-\delta}^{\delta} f_\sigma$ (i.e., the probability of a draw within $\delta$ of the origin), then the opposing player is happy to play if their ante is $(1 - p_0)/p_0$ relative to our one dollar (see logic in [1]).

We similarly find $f_{\sigma_1}$ for the enhanced model that contains side information. We construct $\omega$ based on the single-blind bet using our new side information. Conceptually, if $f$ is good then $\sigma_0^2 < \sigma_0^1$. Our bet of an observation being within $\delta$ is then an increasingly valuable bet in terms of the money we expect to win! We emphasize at this point that $\delta$ is a parameter that needs to be specified. We illustrate this in Figure 1 wherein we consider the IMV associated with different choices of $\delta$ as a function of the true regression parameter $b$ (i.e., $y \sim \mathrm{N}(bx, \sigma^2 = 1)$. Note first that when $b = 0$ the IMV is always 0; information about $x$ offers no value when $x$ is unassociated with $y$. For $b > 0$, the IMV is maximized when $\delta$ is relatively small. [there is intuition here; little value in increased information on very unlikely outcomes. This is due to the fact that the IMV prioritizes increases in uncertainty. There is little value to be had by having additional information about highly certain events.[1]

In the right panel of Figure 1, we also show how we can "equate" different coefficients from linear and logistic regression models via the IMV. For a given value of $\delta$, the curves on the right show the LM and GLM regression parameters that lead to the same IMV values. We choose two values of $\delta$ to illustrate the point; for select points, we also show the IMV values produced by the parameters in the two settings.

[1] Similarly, the IMV is smaller for binary outcomes with prevalences further from 0.5, see [1].

Figure 1: The $\delta$ approach to computing IMV for continuous outcomes. Left: Illustration of IMV as a function of $\delta$ and $b$. Gray vertical lines represent values of $\delta$ used in the right panel. Right: Curves showing values for parameters in linear and logistic regression models that produce the same IMVs for different values of $\delta$.



## 2.2 The expected IMV induced via a dichotomization of the outcome

While there were advantages to the $\delta$-based approach considered above, the ambiguity associated with identification of a $\delta$ may make it suboptimal for many uses. We now consider an alternative strategy. In particular, we consider the following. First, for a value $y^\star$, we discretize the outcome $y$ such that $y = 1$ if $y > y^\star$ and 0 otherwise. We then compute the IMV associated with this discretization. If we denote this as $\omega^\star$, we then compute $\mathbb{E}_{f_y}(\omega^\star)$.

In Figure 2 we illustrate the utility of this approach via simulation. For $x_i \sim \mathrm{N}(0,1)$, we simulate outcomes via $y_i \sim \mathrm{N}(\beta x_i, 1)$ ($i \in \{1,\ldots,5000\}$). At left, we show the IMV as a function of $y^\star$ (x-axis) and $\beta$ (color). Note that the IMV is maximal when $y^\star \approx 0$ which is to be expected given the relationship between prevalence and the IMV previously discussed [1]. At right, we consider the IMV value that results from each choice of $\beta$ via calculation of $\mathbb{E}_{f_y}(\omega^\star)$. The correspondence between $r^2$ and IMV values suggests that we can identify an IMV for predictions of continuous outcomes based on this approach and knowledge of the relevant $r^2$. We will make use of this fact below in the empirical illustrations.

## 2.3 Summarizing the two approaches

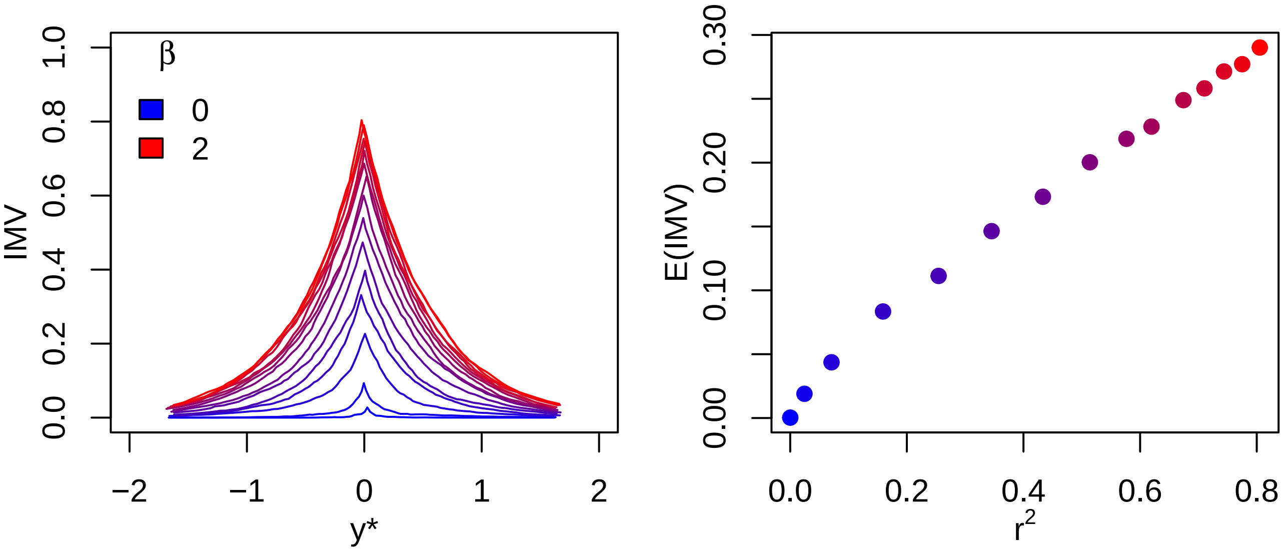
**what did we get from the above two sections? need to offer a blueprint for what we are going to do via simulation in the next section**

# 3 Comparing predictions of binary and continuous outcomes

We pause to emphasize the potential utility of straightforward comparisons between IMV values for various outcome types. Being able to make such straightforward comparisons has, in our view, hampered progress in our general understanding of the degree to which different kinds of outcomes can be readily predicted via existing predictive technologies. Consider the results of [3] which compares the predictability of six lifecourse outcomes; three of these outcomes are continuous while three are dichotomous. The study attempts to straightforwardly interpret the predictability of outcomes despite this fact. We reconsider this analysis in our section containing empirical illustrations.

a coefficient of $\beta_1$ in the logistic regression context here is equivalent to bets involving $\delta$ in the linear regression context. [show comparability using simulation]

Figure 2: IMV via dichotomization. Right: IMV values for different choices of $y^\star$. Left: $\mathbb{E}$(IMV) as a function of $r^2$ for a variety of $\beta$ values.



[show the N needed to detect interactions that have similar IMVs across linear/logistic regressions]

# 4 Empirical Examples

# 5 Discussion

perhaps propose meta-analyses not of effects/coefficient estimates but of IMV values?

## Acknowledgements

This work was supported in part by the Jacobs Foundation.

## References

[1] Ben Domingue, Charles Rahal, Jessica Faul, Jeremy Freese, Klint Kanopka, Alexandros Rigos, Ben Stenhaug, Jialu Streeter, and Ajay Tripathi. Intermodel vigorish (imv): A novel approach for quantifying predictive accuracy when outcomes are binary. 2021.

[2] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[3] Matthew J Salganik, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, Jennie E Brand, Nicole Bohme Carnegie, Ryan James Compton, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020.