# A Extended Discussion

**The scope of the explanation** Model indeterminacy is relevant to the robustness of an explanation when we consider the scope of the explanation we are trying to provide. Regardless of exactly how they frame and produce an explanation, most existing methods operate on a single, fixed predictive model. They explain why a given *model instance*, i.e., a specified model-class with specified parameter values, made a prediction. This is depicted by the inner-most region in Figure 2. However, if there is model indeterminacy, there may have been many approximately equivalent model instances that could have been used in the automated decision system. These could each bring about a different explanation, thus it may be misleading to present an explanation to a user as though the model instance is fully determined. This perspective is supported by the "non-misleading explanations" arguments made by Hancox-Li [11].

In part, the problem is that what precisely constitutes a "model" is ambiguous, especially in communications with an end user. The term model may refer to a specific predictive function with a specific set of parameters (a model instance), just as it may refer to the high-level process of collecting data and making predictions from it. It may equally be used at all levels of abstraction in between. As a result, any explanation of a model's prediction inherits this ambiguity unless we are explicit about the scope of the explanation, which may not be clear to a user. We postulate that users may want—or even expect—explanations which apply to more than just a single model instance, especially if there is no guarantee that the model instance won't just be substituted for an approximately equivalent one at a later time. Explanations which apply to a set of approximately equivalent models need to be robust to model indeterminacy. This would require explanations that are robust to arbitrary choices during model selection (the Rashomon effect), or at least to the intentional use of randomness used during optimization which could affect the specific model instance (underspecification). If we do not at least have robustness to the latter, then the best (local) explanation for a user's prediction may actually be the random seed. In the sense that the model's prediction may fluctuate more due to changes in the random seed than it would due to any small changes in the user's input features.

# B Details of experimentation

## B.1 Dataset details

The datasets are all available on Kaggle. They permit usage for scientific research.

**UCI Credit Card**

- Task: Next month payment prediction (y=1 implies missed payment)
- Details: binary label, 23 features (32 after one-hot-encoding), 30k instances (users)
- Feature Types: 1 binary, 2 categorical, 7 ordinal, 13 continuous
- Immutable features: 3 (age, sex, marriage)

**Give Me Some Credit**

- Task: Predict credit delinquency (y=1 implies serious delinquency)
- Details: binary label, 10 features, over 120k instances (users)
- Feature Types: 7 ordinal, 3 continuous
- Immutable features: 2 (age, number of dependants)

**Porto Seguro's Safe Driver Prediction**

- Task: Predict whether driver will file a claim (y=1 implies claim filed)
- Details: binary label, 34 features (101 after one-hot-encoding), over 595k instances (users)
- Feature Types: 11 binary, 13 categoricals, 4 ordinal, 6 continuous
- Immutable features: 2 (regional features), others are likely but not indicated

### B.2 Model sweep, best models and hyperparameters

Here we include details of the model and hyperparameter sweep, as well as the best models and hyperparameter configurations found. We tabulate the results of the sweep and formation of approximately equivalent model sets in Table 2. We report the max and 97th percentile validation AUROCs from the sweep, which bound the validation performance of the model instances in the Rashomon effect sets. We also report the model class(es) which compose each type of set, as well summarize the performance of the models in each set by indicating the min, mean, and max holdout AUROC.

The Rashomon effect sets used the top 3% of models instances, while the underspecification sets used 100 model instances trained (with different random seeds) from the the best model and hyperparameter settings for each dataset.

Table 2: Hyperparameter sweep and set formation

| | Sweep | | Rashomon effect | | | | Underspecification | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUROC (val.) | | classes | AUROC (holdout) | | | class | AUROC (holdout) | | |
| | 97th | max | | min | mean | max | | min | mean | max |
| UCI Credit Card | 0.778 | 0.780 | 12 MLP, 24 TRN | 0.762 | 0.770 | 0.775 | TRN | 0.771 | 0.773 | 0.775 |
| Give Me Credit | 0.832 | 0.857 | 36 MLP | 0.831 | 0.844 | 0.852 | MLP | 0.849 | 0.853 | 0.856 |
| P.S.'s Safe Driver | 0.636 | 0.639 | 25 MLP, 11 TNR | 0.630 | 0.633 | 0.635 | MLP | 0.635 | 0.636 | 0.638 |

**UCI Credit Card (best model & hyperparameters)**

- model=tabresnet
- model.args.n_blocks=2
- model.args.d_main=128
- model.args.d_hidden=256
- model.args.dropout_first=0.25
- model.args.dropout_second=0
- train.initial_lr=0.01
- train.weight_decay=0.001
- train.batch_size=4096
- train.reduce_lr_by=0.9
- train.num_epochs=71

**Give Me Some Credit (best model & hyperparameters)**

- model=mlp
- model.args.hidden_layer_sizes=[64]
- train.initial_lr=0.1
- train.weight_decay=0.0
- train.batch_size=4096
- train.reduce_lr_by=0.99
- train.patience=500
- train.num_epochs=324

**Porto Seguro's Safe Driver Prediction (best model & hyperparameters)**

- model=mlp
- model.args.hidden_layer_sizes=[2056]
- train.initial_lr=0.1

- train.weight_decay=0.001
- train.batch_size=4096
- train.reduce_lr_by=0.95
- train.patience=500
- train.num_epochs=301

### B.3 Rashomon Sets vs. Rashomon *Effect* Sets

Semenova et al. [30] define the empirical Rashomon set (or simply Rashomon set) as the set of all models in the hypothesis space having empirical risk within some small epsilon > 0 of the optimal risk (achieved by the empirical risk minimizer). We use this definition in our linear analysis.

What we call "Rashomon effect sets" in our experiments differ in three ways:

1. We used the AUROC on the validation sets to determine which models were in the Rashomon set, not the empirical risk. We felt this was closer to what would be used in a practical model selection process.
2. We consider only a finite sample of models, those which were found during a realistic model sweep. We do not consider all models in the hypothesis space as it is impractical in our settings (neural nets or other complex hypothesis spaces).
3. While the development data is fixed during model selection, the train / validation split is controlled by a random seed (which takes one of three values). Therefore, the data used for training v.s. stopping may vary from one instance to the next.

## C Details of linear analysis

Here we provide the details of the derivations supporting the claims discussed in the linear analysis in Section 3. Recall we are considering the case where $\boldsymbol{x} \in \mathbb{R}^M, y \in \mathbb{R}$, and $\mathcal{F}$ is the set of ridge regression models of the form $f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{x}$, with square error loss and $\ell_2$ penalty $\alpha > 0$. We are interested in circumstances whereby the sign of a Shapley value reverses within this Rashomon set, i.e., $\exists\, \boldsymbol{\theta}, \boldsymbol{\theta'} \in \Theta_\varepsilon$ *s.t.* $\phi_k(\boldsymbol{x}; \boldsymbol{\theta}) > 0 > \phi_k(\boldsymbol{x}; \boldsymbol{\theta'})$.

*Note: For illustrative simplicity we assume that* $\sum_n \boldsymbol{x}_n = \mathbf{0}$ *and* $\sum_n y_n = 0$ *throughout.*

**Claim** For a linear model of the form $f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^M$, using reference distribution $R \sim \mathcal{D}_{\text{ref}}$, with $\mathbb{E}[R] = \bar{\boldsymbol{x}}$, the Shapley value of feature $i$ is $\phi_i(\boldsymbol{x}; f) = \theta_i(x_i - \bar{x}_i)$.

**Derivation** Recall to Equation 1. The Shapley value for feature $i$ given input $\boldsymbol{x}$ can then be written as

$$\phi_i(\boldsymbol{x}; f) = \frac{1}{M} \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{i\}} \binom{M-1}{|\mathcal{S}|}^{-1} \mathbb{E}_{R \sim \mathcal{D}_{\text{ref}}} \left[ f\left(\boldsymbol{u}\left(\boldsymbol{x}, R, \mathcal{S} \cup \{i\}\right)\right) - f\left(\boldsymbol{u}\left(\boldsymbol{x}, R, \mathcal{S}\right)\right) \right].$$

Where $\mathcal{M}$ is the set of input features, with $|\mathcal{M}| = M$, $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{r}, \mathcal{S})$ is a composite input function, which takes the values of $\boldsymbol{x}$ for all the features in $\mathcal{S} \subset \mathcal{M}$, and otherwise takes the values of $\boldsymbol{r}$, some other reference input. The sum is over all possible subsets of features (excluding feature $i$). Each term is the marginal change in the prediction when including feature $i$, divided by the binomial coefficient $M - 1$ choose $|S|$, and $\mathcal{D}_{\text{ref}}$ is the reference distribution.

When $f$ is linear we have

$$\begin{aligned} f(\boldsymbol{u}(\boldsymbol{x}, R, S \cup \{i\})) - f(\boldsymbol{u}(\boldsymbol{x}, R, S)) &= \boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x}, R, S \cup \{i\}) - \boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x}, R, S) \\ &= \boldsymbol{\theta}^T \big(\boldsymbol{u}(\boldsymbol{x}, R, S \cup \{i\}) - \boldsymbol{u}(\boldsymbol{x}, R, S)\big) \\ &= \theta_i(x_i - r_i), \end{aligned}$$

where $r_i$ denotes the $i^{\text{th}}$ component of $R$. The last line follows from the fact that $\boldsymbol{u}(\boldsymbol{x}, R, S \cup \{i\})$ and $\boldsymbol{u}(\boldsymbol{x}, R, S)$ differ only in the $i^{\text{th}}$ component. Equation 1 can be rewritten because the only term

varying in the sum is $|S|$,

$$\phi_i(\boldsymbol{x}; f) = \mathbb{E}_{R\sim\mathcal{D}_{\text{ref}}}[\theta_i(x_i - r_i)]\left[\frac{1}{M}\sum_{S\subseteq\mathcal{M}\setminus\{i\}}\binom{M-1}{|S|}^{-1}\right].$$

The sum is over the superset of $\mathcal{M}\setminus\{i\}$, so there are $2^{M-1}$ terms in the sum. However, $|S|$ only takes on $M$ different values: 0, 1, 2, ... $M-1$. Moreover, $|S| = k$ exactly $\binom{M-1}{k}$ times. Therefore we have,

$$\begin{aligned}
\phi_i(\boldsymbol{x}; f) &= \mathbb{E}_{R\sim\mathcal{D}_{\text{ref}}}[\theta_i(x_i - r_i)]\left[\frac{1}{M}\sum_{k=0}^{M-1}\binom{M-1}{k}\binom{M-1}{k}^{-1}\right] \\
&= \mathbb{E}_{R\sim\mathcal{D}_{\text{ref}}}[\theta_i(x_i - r_i)]\left[\frac{1}{M}\sum_{k=0}^{M-1}1\right] \\
&= \mathbb{E}_{R\sim\mathcal{D}_{\text{ref}}}[\theta_i(x_i - r_i)] \\
&= \theta_i(x_i - \overline{x}_i).
\end{aligned}$$

If $\overline{\boldsymbol{x}} = \mathbf{0}$, which is to say the reference distribution also has a zero mean, then $\phi_i(\boldsymbol{x}; f) = \theta_i x_i$

**Claim** Given the set of ridge regression models of the form $f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^T\boldsymbol{x}$, $\boldsymbol{x}\in\mathbb{R}^M$, with $\ell_2$ loss, $L_{\boldsymbol{\theta}} = \sum_{n=1}^N (y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n)^2 + \alpha\|\boldsymbol{\theta}\|^2$, $\alpha > 0$. The Rashomon set of all $\boldsymbol{\theta}$ which parameterize models within $\varepsilon$ of the optimal loss, $L_{\boldsymbol{\theta}^*}$, is an ellipsoidal region defined by $\Theta_\varepsilon = \{\boldsymbol{\theta}\in\mathbb{R}^M : (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\le\varepsilon\}$. Where $\boldsymbol{\theta}^*$ is the OLS solution, and $\boldsymbol{H} = [\sum_n \boldsymbol{x}_n\boldsymbol{x}_n^T + \alpha\boldsymbol{I}] \succ 0$.

**Derivation** Define the $N$ by $M$ data matrix of observed features as $\boldsymbol{X}$, and the vector of $N$ labels as $\boldsymbol{y}$. With this notation, the loss can then be written as,

$$\begin{aligned}
L_{\boldsymbol{\theta}} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \alpha\boldsymbol{\theta}^T\boldsymbol{\theta} \\
&= \boldsymbol{\theta}^T H\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}
\end{aligned}$$

The OLS solution is $\boldsymbol{\theta}^* = \boldsymbol{H}^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where we can re-write $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I}$. (For proof, see an introductory ML text such as Murphy [25]) We can then write the difference from the optimal loss as:

$$\begin{aligned}
L_{\boldsymbol{\theta}} - L_{\boldsymbol{\theta}^*} &= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* + 2\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{y} \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* + 2\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{y} \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T[\boldsymbol{H}\boldsymbol{H}^{-1}]\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* + 2\boldsymbol{\theta}^{*T}[\boldsymbol{H}\boldsymbol{H}^{-1}]\boldsymbol{X}^T\boldsymbol{y} \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{H}[\boldsymbol{H}^{-1}\boldsymbol{X}^T\boldsymbol{y}] - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* + 2\boldsymbol{\theta}^{*T}\boldsymbol{H}[\boldsymbol{H}^{-1}\boldsymbol{X}^T\boldsymbol{y}] \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta}^* - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* + 2\boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* \\
&= \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} - \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*T}\boldsymbol{H}\boldsymbol{\theta}^* \quad \textit{(since } \boldsymbol{H} \textit{ is symmetric)} \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)
\end{aligned}$$

**Claim** If the features of $\boldsymbol{x}$ are independent, in the sense that we assume $\boldsymbol{X}^T\boldsymbol{X}$ is diagonal, then for $\boldsymbol{\theta}\in\Theta_\varepsilon$, we have $\theta_k$ bounded in an interval centered at $\theta_k^*$, having width $2(\varepsilon/(\sum_{n=1}^N x_{k,n}^2 + \alpha))^{1/2}$. *(For illustrative simplicity we are assuming a form for $\boldsymbol{X}^T\boldsymbol{X}$ that is only the case in the limit as $N$ goes to infinity.)*

**Derivation** From the claim above we have $\Theta_\varepsilon = \{\boldsymbol{\theta}\in\mathbb{R}^M : (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\le\varepsilon\}$. If $\boldsymbol{X}^T\boldsymbol{X}$ is diagonal, then so is $\boldsymbol{H}$, with the $k^{\text{th}}$ entry being $\lambda_k = \sum_n x_{k,n}^2 + \alpha$. Therefore, the ellipsoidal region is axis-aligned, and defined by

$$\sum_{i=1}^M \lambda_i(\theta_i - \theta_i^*)^2 \le \varepsilon.$$

Since all terms in the sums are positive, the extrema of $\theta_k$ occur when $\theta_i = \theta_i^* \; \forall i \neq k$. Those extrema occur when $\lambda_k(\theta_k - \theta_k^*)^2 = \varepsilon$, thus at

$$\begin{aligned}\theta_k &= \theta_k^* \pm \sqrt{\frac{\varepsilon}{\lambda_k}} \\ &= \theta_k^* \pm \sqrt{\frac{\varepsilon}{\sum_n x_{k,n}^2 + \alpha}}.\end{aligned}$$

**Claim** If all features of $\boldsymbol{x}$ are independent other than features $j$ and $k$, which have variance $\sigma^2$ and $\mathrm{Cov}(x_j, x_k) = \beta$, $0 < \beta < \sigma^2$, in the sense that we assume $\boldsymbol{X}^T\boldsymbol{X}$ is diagonal other than for two additional non-zero elements $[\boldsymbol{X}^T\boldsymbol{X}]_{j,k} = [\boldsymbol{X}^T\boldsymbol{X}]_{k,j} = \beta$, and that $[\boldsymbol{X}^T\boldsymbol{X}]_{j,j} = [\boldsymbol{X}^T\boldsymbol{X}]_{k,k} = \sigma^2$, then we have $\theta_k$ bounded in an interval centered at $\theta_k^*$. If we fix $\varepsilon$ and $\sigma^2$, the width of the interval grows with $\beta$ but shrinks with $\alpha$. *(For illustrative simplicity we are assuming a form for $\boldsymbol{X}^T\boldsymbol{X}$ that is only the case in the limit as $N$ goes to infinity.)*

**Derivation** The ellipsoidal region $\Theta_\varepsilon$ is axis-aligned for all features other than $j$ and $k$. It is defined by

$$(\sigma^2 + \alpha)(\theta_j - \theta_j^*)^2 + 2\beta(\theta_j - \theta_j^*)(\theta_k - \theta_k^*) + (\sigma^2 + \alpha)(\theta_k - \theta_k^*)^2 + \sum_{i \notin \{j,k\}} \lambda_i(\theta_i - \theta_i^*)^2 \leq \varepsilon.$$

Where $\lambda_k = \sum_n x_{k,n}^2 + \alpha$. All terms in the sum over $i$ are positive, thus the extrema of $\theta_k$ occur when $\theta_i = \theta_i^* \; \forall i \notin \{j, k\}$, when

$$\sigma^2(\theta_j - \theta_j^*)^2 + 2\beta(\theta_j - \theta_j^*)(\theta_k - \theta_k^*) + \sigma^2(\theta_k - \theta_k^*)^2 = \varepsilon.$$

We can parameterize this bounding ellipse as

$$\begin{aligned}\theta_j(t) &= \theta_j^* + \sqrt{\frac{\varepsilon}{(\sigma^2 + \alpha + \beta)}}\cos(\pi/4)\cos(t) - \sqrt{\frac{\varepsilon}{2(\sigma^2 + \alpha - \beta)}}\sin(\pi/4)\sin(t) \\ \theta_k(t) &= \theta_k^* + \sqrt{\frac{\varepsilon}{(\sigma^2 + \alpha + \beta)}}\sin(\pi/4)\cos(t) + \sqrt{\frac{\varepsilon}{(\sigma^2 + \alpha - \beta)}}\cos(\pi/4)\sin(t).\end{aligned}$$

It is centered at $(\theta_j^*, \theta_k^*)$, rotated $\pi/4$ off axis, and the length of the major axis is

$$2\sqrt{\frac{\varepsilon}{(\sigma^2 + \alpha - \beta)}}.$$

For a fixed $\varepsilon$ and $\sigma^2$ we see that the length of the major axis, and thus the range of $\theta_k$, increases with the covariance, $\beta$, and decreases with the $\ell_2$ penalty, $\alpha$.

# D Additional experimental results

## D.1 Consistency of decisions

Table 3: Consistency of decisions

| **Dataset** | UCI Credit Card | | | Give Me Some Credit | | | P.S.'s Safe Driver | | |
|---|---|---|---|---|---|---|---|---|---|
| users accepted at risk threshold | 25% | 50% | 75% | 25% | 50% | 75% | 25% | 50% | 75% |
| **Rashomon effect** | | | | | | | | | |
| probability of "bad" outcome (%) | 6.93 | 9.86 | 12.81 | 0.89 | 1.36 | 2.30 | 1.86 | 2.32 | 2.84 |
| probability of a decision flip (%) | 7.16 | 7.98 | 3.50 | 11.79 | 12.7 | 8.05 | 9.39 | 11.59 | 8.66 |
| **Underspecification** | | | | | | | | | |
| probability of "bad" outcome (%) | 6.73 | 9.59 | 12.91 | 0.75 | 1.19 | 2.19 | 1.83 | 2.27 | 2.84 |
| probability of a decision flip (%) | 3.67 | 4.35 | 1.96 | 6.95 | 6.32 | 4.40 | 5.11 | 6.25 | 4.99 |

Here we include an analysis of the consistency of the decisions that the users may receive. We are cognisant that in a risk assessment setting, the final decisions will be a function of an institution's risk appetite. Therefore, rather than assuming the risk threshold (above which users are denied and below which users are accepted) we consider a range of thresholds. Each threshold corresponds to a user acceptance rate: the percentage of users that will be accepted at this risk level. We consider three quantities of interest as a function of this acceptance rate. First, the false omission rate, which is the probability that an accepted user will have a "bad" outcome, i.e., missed payment, serious delinquency, or file an insurance claim. This could be used together with the acceptance rate, and other cost/revenue information to pick a decision threshold. We then consider the pairwise inter-model decision agreement for each user in the holdout set. The agreement is 100% if two model make the same decision for every user, and 0% if they make opposite decisions for every user. Intuitively, if models are thresholded so as to either accept every user, or reject every user, they will all show 100% agreement. Finally, we consider the probability that a decision given to a user flip from accept to reject, or from reject to accept, if we were to switch to approximately equivalent model. The results are tabulated in Table 3 for three acceptance levels. They are plotted over 5% to 95% acceptance levels for all datasets in Figure 6.

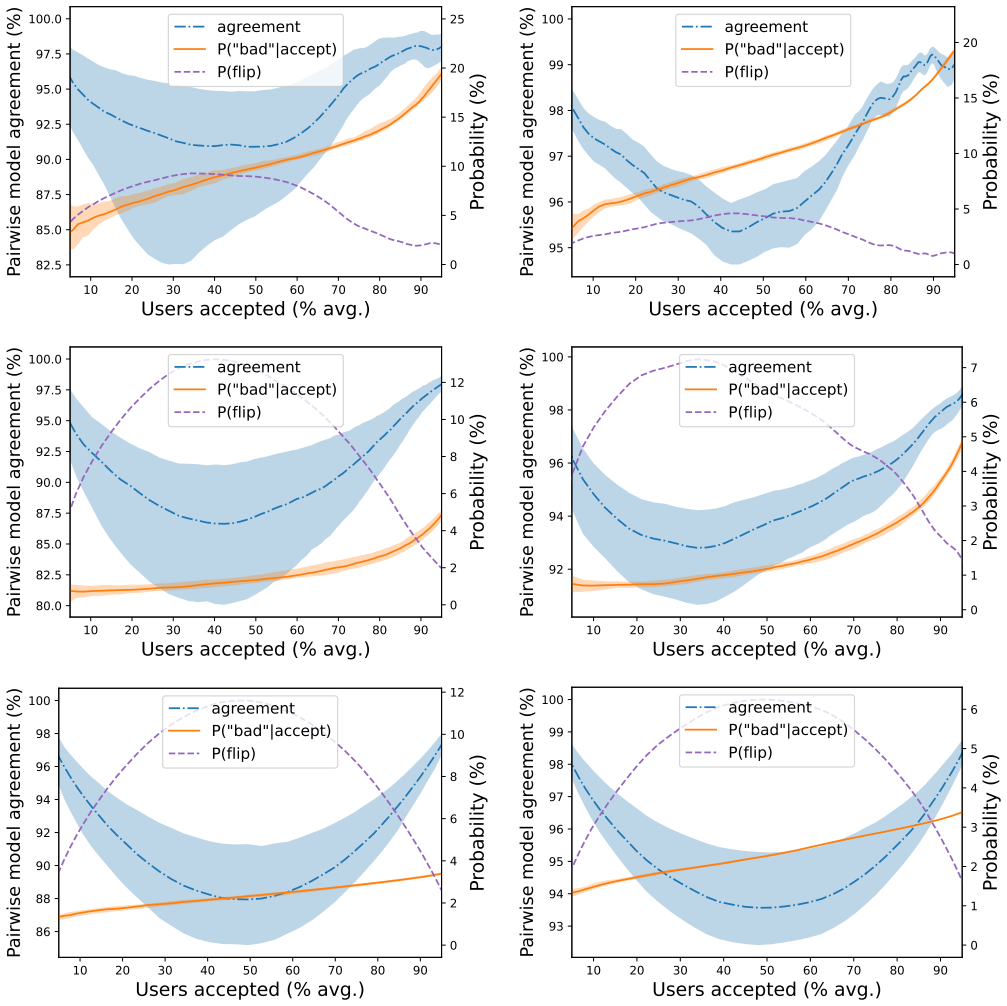


Figure 6: We vary the risk threshold and observe its effect on the the consistency of decisions. On the horizontal axis we show the user acceptance rate for a given risk threshold, ranging from 5% to 95%. This is the average rate over approximately equivalent models. At each acceptance rate we plot the corresponding false omission rate, $p(\text{“bad”}|\text{accept})$, i.e., the probability that an accepted user will nonetheless have have a “bad” outcome. We also plot the pairwise model agreement, the percent of decisions that the models agree upon. For each of these, we trace the median value, and shade the region between the 5th and 95th percentiles. Finally, we plot the probability that a user's decision would flip if we switched between two approximately equivalent models, $P(\text{flip})$. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction Columns (left, right): Rashomon effect, Underspecification]

### D.2 Probability of Disagreement (SSD)

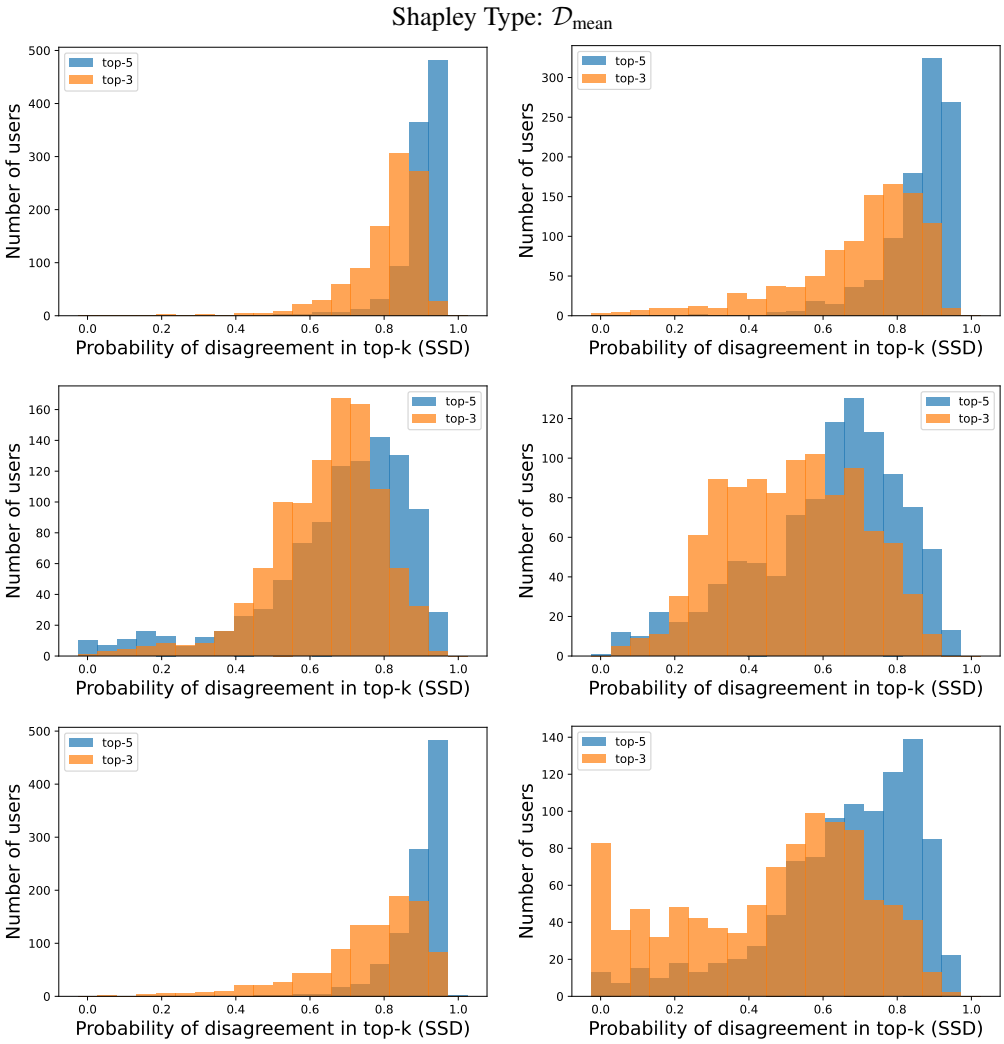


Figure 7: We compare the top-k Shapley values across approximately equivalent models for 1000 users in the hold-out set. We show histograms over the probability of inter-model explanation disagreement for each user. Agreement entails that two models have the same k highest magnitude Shapley values. The sign of their values must agree, but the relative order within the top-k need not. (See SSD metric in Section 4.1). [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

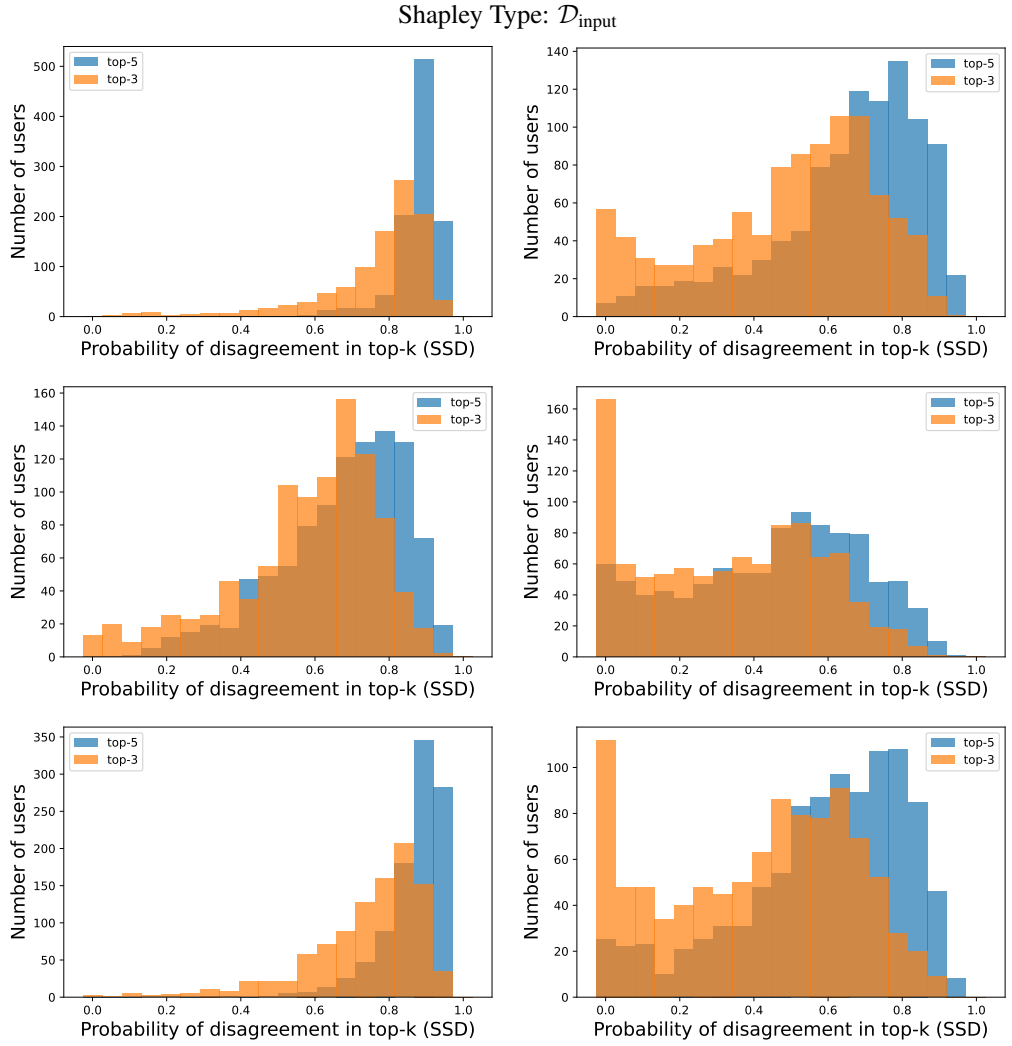


Figure 8: Same as Figure 7 but for $\mathcal{D}_{\text{input}}$. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

### D.3 Probability of Contradiction (CDC)

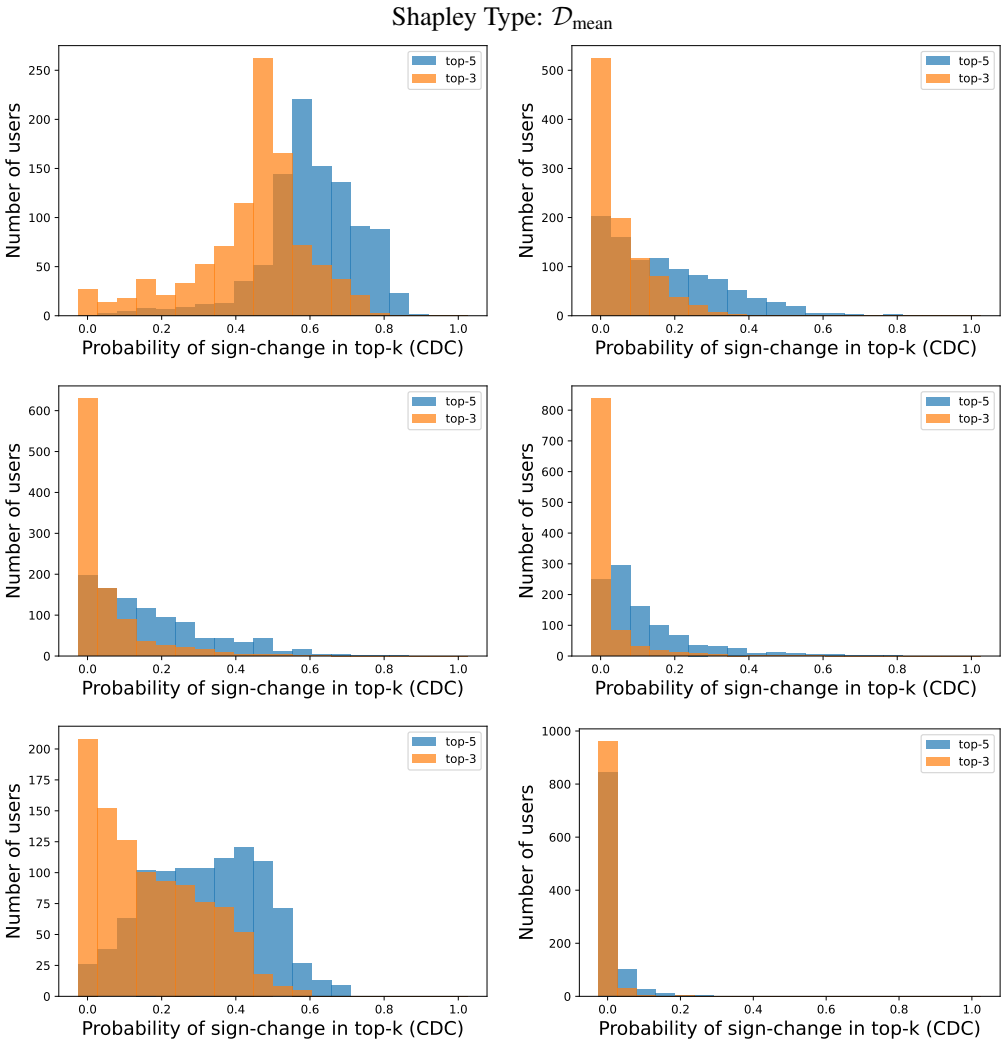


Figure 9: Contradictions. We compare the top-k Shapley values across approximately equivalent models for 1000 users in the hold-out set. We show histograms over the probability of inter-model contradiction (sign-change) in the top-k Shapley values for each user. (See CDC metric in Section 4.1). [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

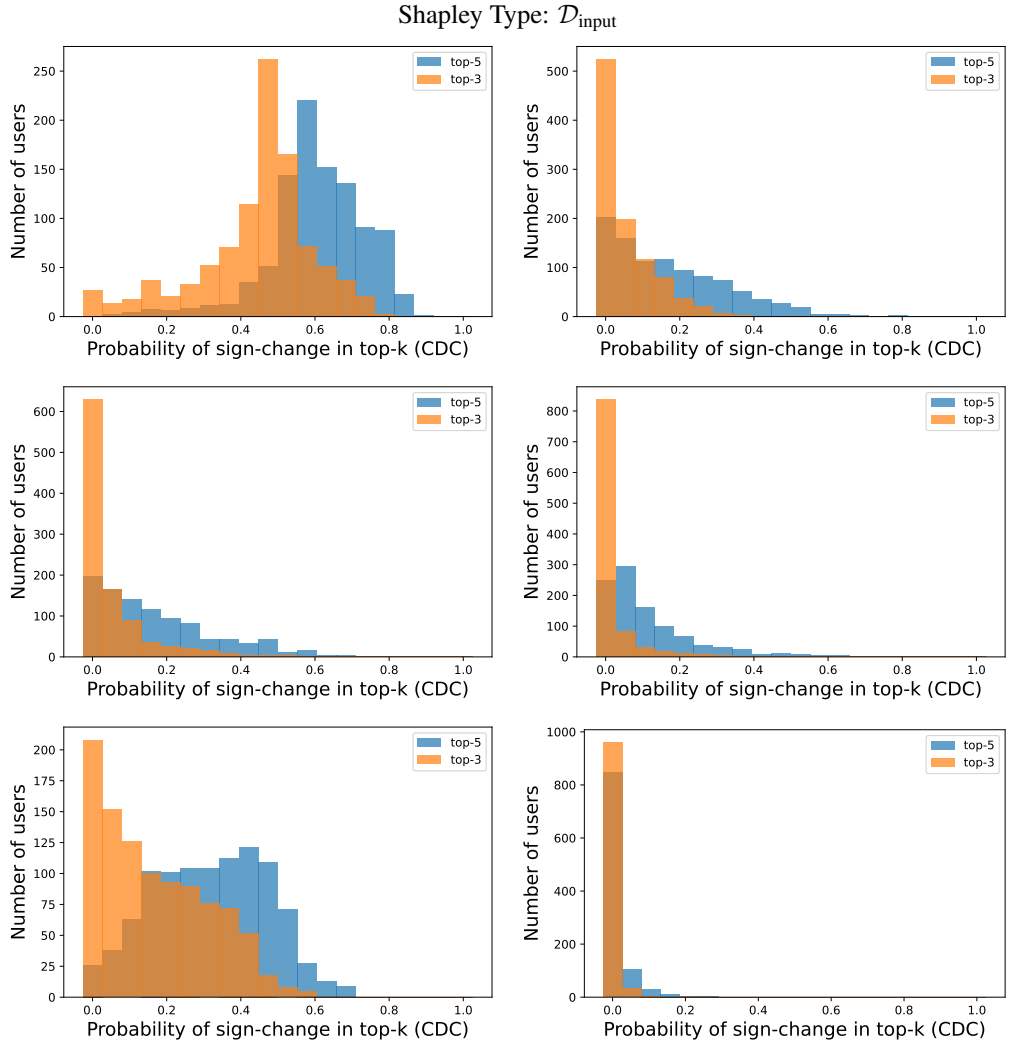


Figure 10: Same as Figure 9 but for $\mathcal{D}_{\text{input}}$. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

### D.4 Visualizing inconsistencies in explanations

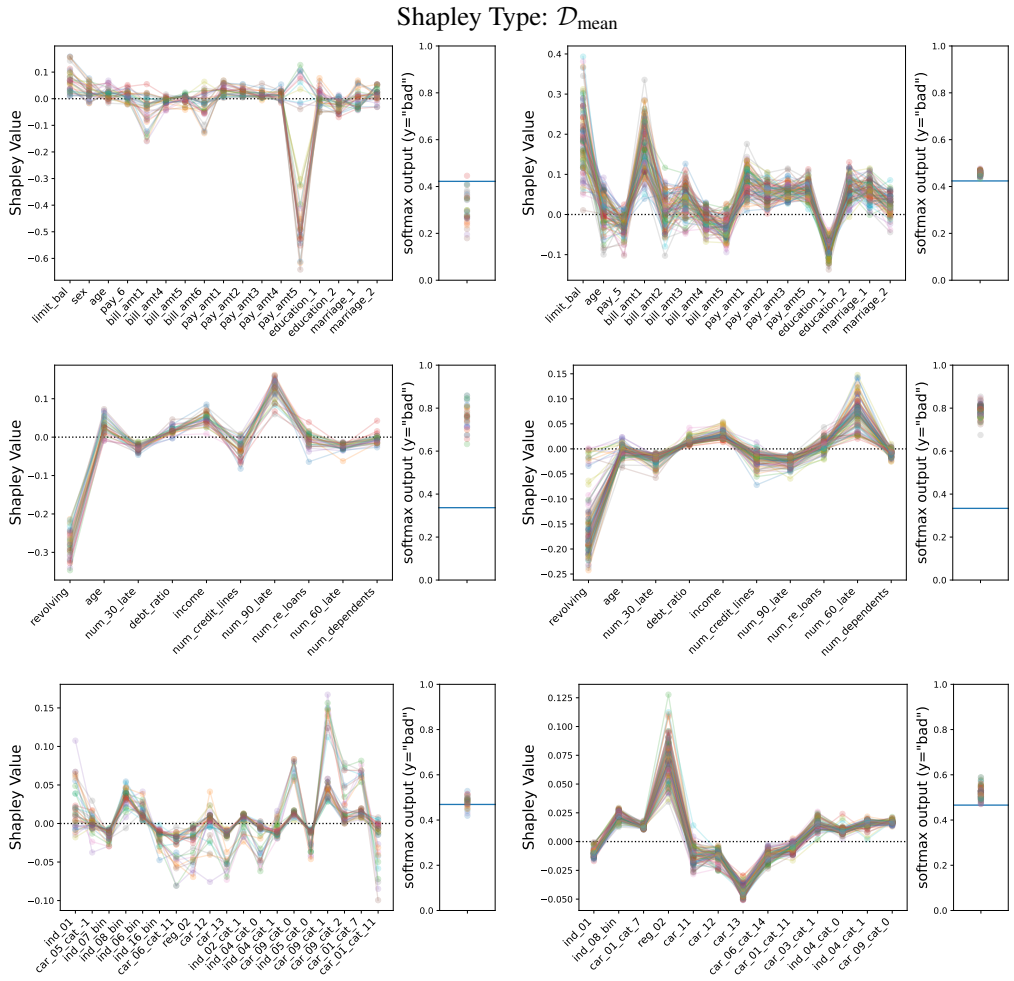


Figure 11: We sort users by the sign agreement (SA) in their top-5 Shapley values across approximately equivalent models. We select the user at the 10th percentile, where a lower percentile implies less agreement. On the horizontal axis we show the union of the top-5 features across all models. On the vertical axis we show the Shapley value for that feature. A positive value indicates the feature is adding to the model's predicted risk, a negative value indicates the feature is subtracting from the predicted risk. Each trace corresponds to one model's explanation. Notice that the Shapley value for some features cross the 0 line. This means that for some models they would be presented among the top-5 reasons that a decision was made, yet for other approximately equivalent models, they would be explained as having the *opposite* effect on that decision. On the right, we show the models' softmax outputs which would be thresholded (based on risk appetite) to make a decision. Notice in some cases they are tightly clustered yet the explanations are very dissimilar. In these cases, the models would tend to agree in their decisions, but not in their explanations. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

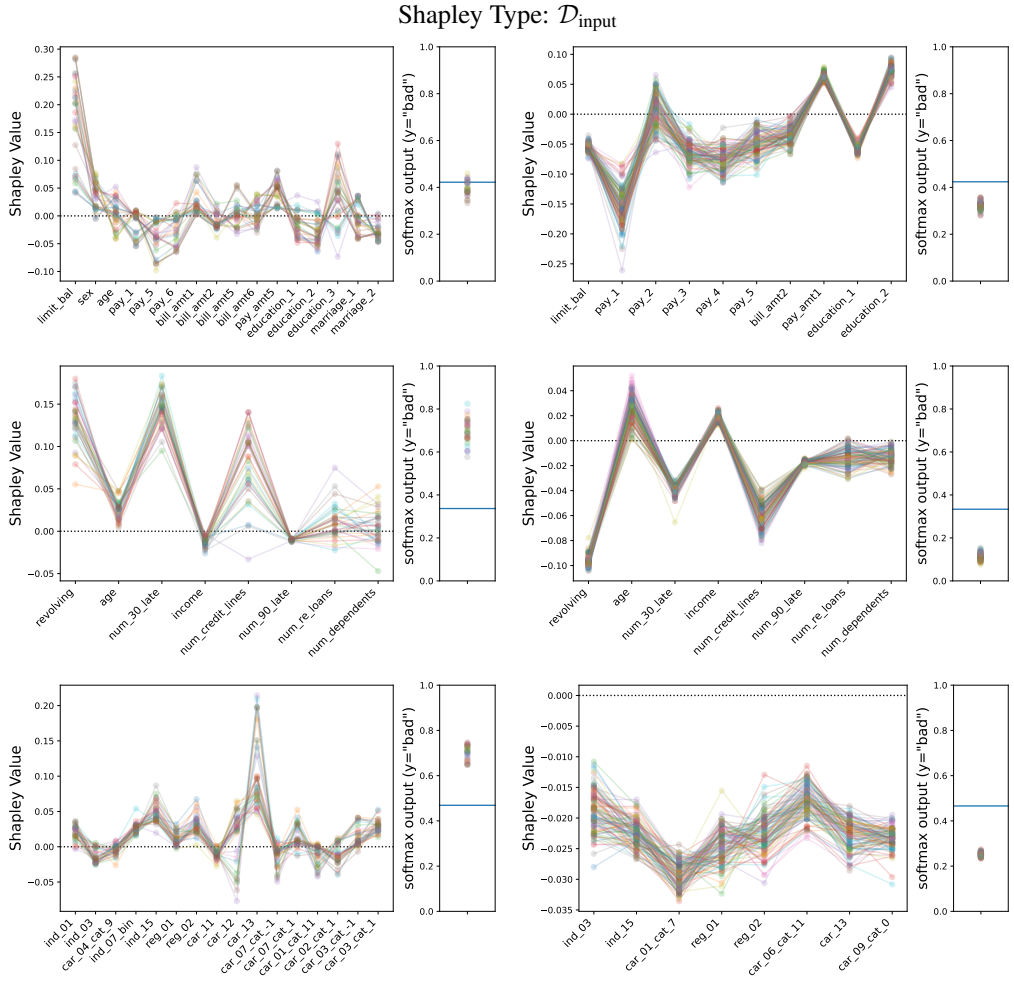


Figure 12: Same as for Figure 11 but for $\mathcal{D}_{\text{input}}$. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

### D.5 Predictive vs. explanatory multiplicity

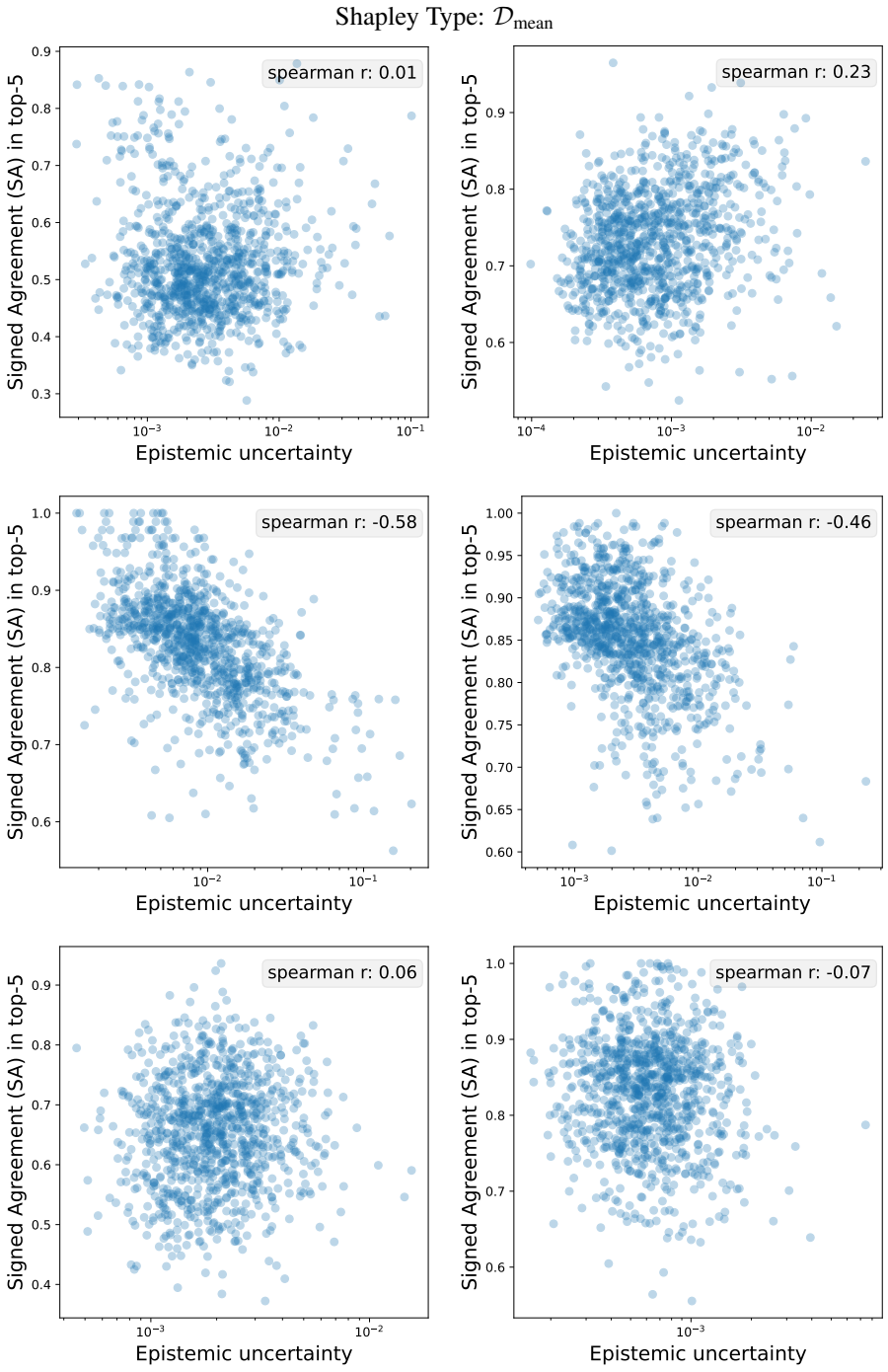


Figure 13: We illustrate the relationship between predictive multiplicity and explanatory multiplicity. We plot the sign agreement (SA) in the top-5 Shapley values vs. the (log) epistemic uncertainty in the predictions for 1000 users in the holdout set. We report the Spearman correlation coefficient (r) in each figure. We do not see a consistent correlation. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

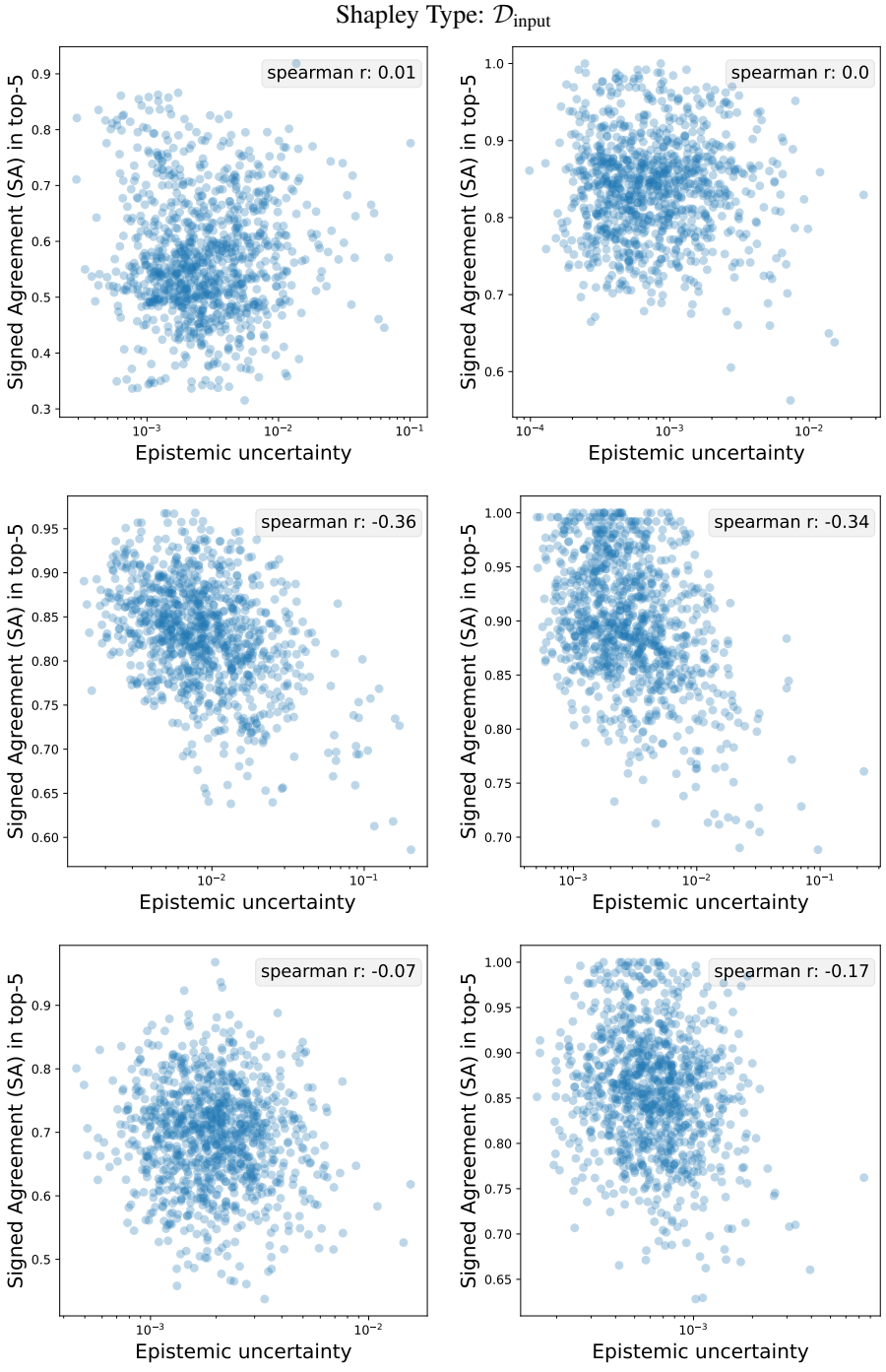


Figure 14: Same as Figure 13 but for $\mathcal{D}_{\text{input}}$. [Rows (top to bottom): UCI Credit Card, Give Me Some Credit, Porto Seguro's Safe Driver Prediction. Columns (left, right): Rashomon effect, Underspecification]

## E Illustrative Example

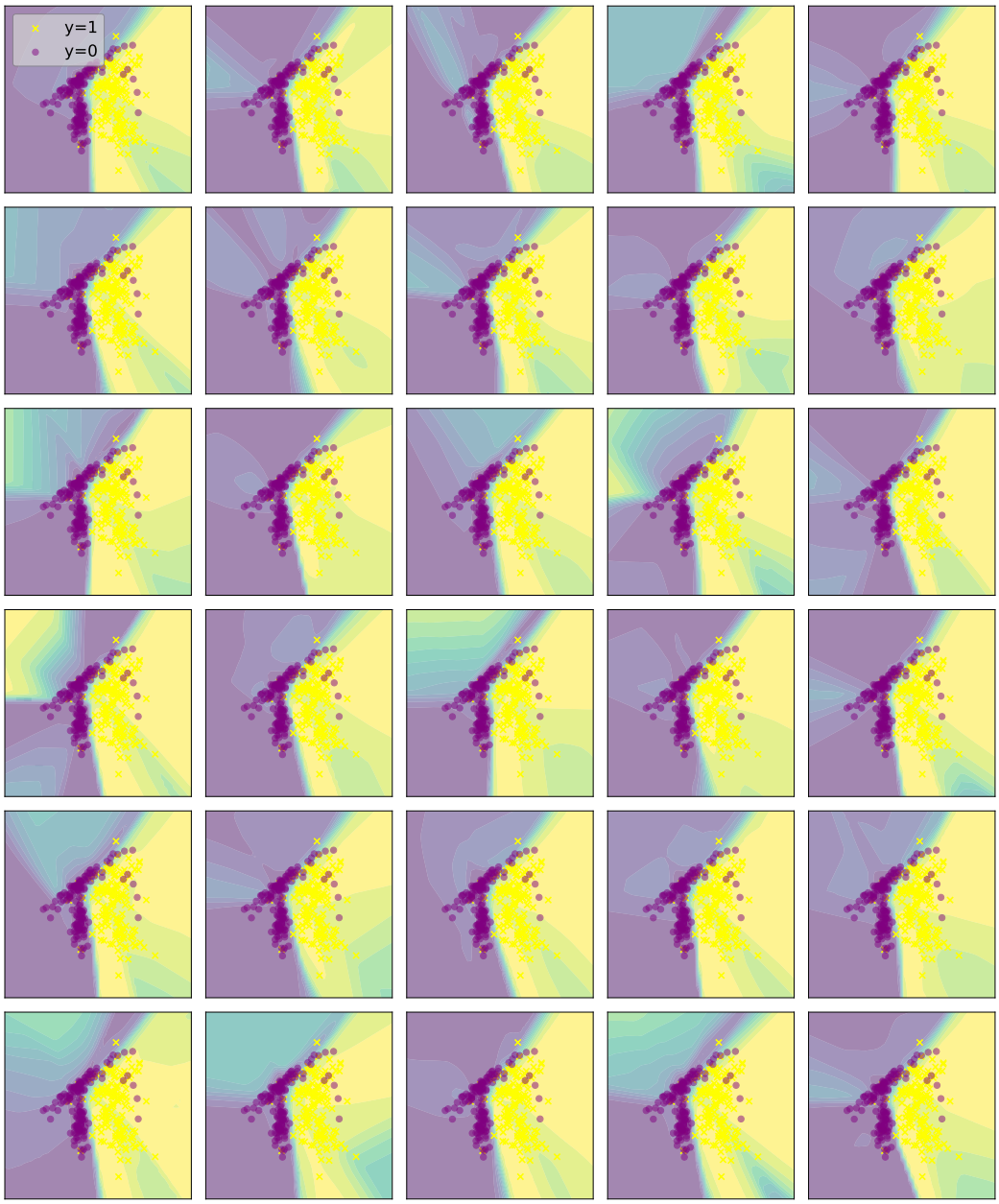


Figure 15: Thirty (30) neural networks, differing only in their random seeds, are trained on the same binary classification dataset. The contour map indicates the model output (softmax probability). Despite showing very similar accuracy (93.3% - 93.6%), we see substantial differences in the form of the learned predictive functions.

## F Shapley Value Estimates

In Equation 1 there is a sum over all subsets of features in $\mathcal{M}$ which grows as $\mathcal{O}(2^{|\mathcal{M}|})$. In practice, this sum must be approximated for inputs with even a modest number of features. We use a simple antithetic method, described in [23], and sample 200 terms from the sum. We found that this was enough for the relative order of the top Shapley values to stabilize. An example of the convergence is depicted in Figure 16. Importantly, we were very careful to ensure that for a given point/user, the explanations across all the models were computed with the exact same set of samples (i.e. the same terms from the sum). While sampling may cause the explanations to deviate slightly from what they would be if we included every term in the sum, all the models are explained using the same set of samples, so it is not what is driving the inter-model explanatory disagreement. If two model instances agreed point-wise in their domains, so would their explanations. Additionally, experiments were run on lower dimensional datasets where every term from the sum in Equation 1 was included. Explanatory multiplicity was very much present in this setting as well.

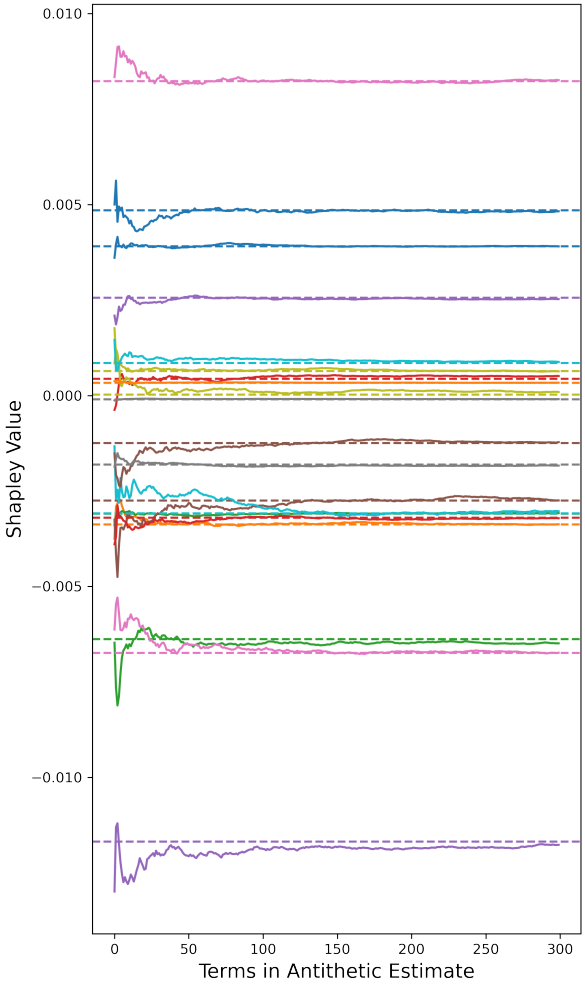


Figure 16: Convergence of Shapley value estimate for a model with M=20 input features. Dotted lines indicate the true Shapley value, which can be still be computed exactly at this dimensionality.