[0191] Model ensembling, which is described by Z.-H. Zhou in the paper titled "Ensemble methods: foundations and algorithms". CRC press, 2012, the entire contents of which are incorporated by reference, is the research of combining multiple pre-trained models to achieve better performance than individual models. A naïve way of ensembling two models is to train a linear regressor to fit scores from the two pre-trained models, which is known as linear stacking. Considering the data instance x and two pre-trained models M_a and M_b, the linear stacking result $LS(x)$ may be computed according to the following Equation (2):

$$LS(x) = w_1 \cdot M_a(x) + w_2 \cdot M_b(x) \tag{2}$$

[0192] Feature-weighted linear stacking (FWLS), which is described by J. Sill, G. Takács, L. Mackey, and D. Lin in the paper titled "Feature-weighted linear stacking," arXiv preprint arXiv:0911.0460, 2009, the entire contents of which are incorporated by reference, is a state-of-the-art model ensembling solution, claiming that the weights (w_1 and w_2 in Equation (2)) should not be fixed but vary based on different feature-values because the two models may have varying performances in different feature value ranges. FWLS, therefore, combines data features with models' scores by feature-crossing and trains the regressor on the crossed features. The weights w_1 and w_2 become m pairs of weights w_1^i (and w_2^i), where m is the number of features and $F_i(x)$ retrieves the ith feature value of x according to the following Equation (3):

$$FWLS(x) = \sum_{i=1}^{m} [w_1^i \cdot (M_a(x) \cdot F_i(x)) + w_2^i \cdot (M_b(x) \cdot F_i(x))] \tag{3}$$

[0193] As the number of features may be large, FWLS can be conducted using a subset of the features. However, incautiously selected features may impair the ensembling result. Therefore, properly ranking the features and choosing the most impactful or important ones for ensembling becomes a problem.

[0194] In some non-limiting embodiments or aspects, learning from disagreement at step 308 of process 300 additionally, or alternatively, includes determining a relative success rate of a first ML model and a second ML model. For example, model comparison system 102 may determine a relative success rate of the first ML model and the second ML model. As an example, model comparison system 102 may

generative adversarial networks (GANs)), where the two compared models explicitly identify where to learn and the discriminator is similar to the discriminator of GAN. This expert also identified the advantage of LFD in “smartly avoiding the data imbalance issue”. E2 has decades of experience in model building and feature engineering. This expert commented “as a model designer, the feature order is extremely important” and appreciated the feature-ordering metrics introduced. This expert also believed that proposing new meta-features is a “good and intuitive” way to “probe the behaviors of the compared models”. The insights from LFD inspired him to revisit early works that filter instances by a model’s score and use the filtered ones to further improve another model. E4-E5 are frontline ML practitioners, working with ML models in production. They especially liked the interactivity of the system in ranking meta-features and the efficacy of LFD in identifying the complementary ones for model ensembling.

[0263] There were also some desired features for LFD and the system. For example, other feature interpretation methods may be integrated into LFD and the framework may be made more general. Negative side comparisons may also be enabled (e.g., comparing true-negative and false-negative instances by sorting instances increasingly at Step 2) to extend LFD. The precision of the two compared models at different thresholds may also be provided in the Disagreement Distribution view.

[0264] Although non-limiting embodiments or aspects of the present disclosure focus primarily on positive predictions (e.g., TP and FP instances) of the two compared classifiers, and thus, the instances are at least captured by one model, they are not limited thereto and two classifiers may be compared from the negative predictions (e.g., sorting the scores increasingly at Step 2 of LFD). It is also noted that the A-B- (as well as the A+B+) cell may be of less interest for the purpose of comparison, because it is where the two classifiers agree on, and there is no disagreement to learn from.

[0265] LFD may depend on SHAP as to the interpretation accuracy. Consequently, LFD may have an inherent limitation on cases where SHAP cannot provide accurate interpretations. For this limitation, two points are noted. First, SHAP is widely applicable to most ML models and comes with solid theoretical supports. Therefore, considerably inaccurate interpretations are not expected to occur very often. Second, as the six steps of LFD have been very well modularized, SHAP may be easily replaced with other interpretation methods at Step 6 (e.g., the recently proposed