Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction

Mohsen Shahhosseini[1], Guiping Hu[2*], and Hieu Pham[3]

[1,2,3] Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, Iowa, 50011, United States

ABSTRACT

Designing ensemble learners has been recognized as one of the significant trends in the field of data knowledge especially in data science competitions. Building models that are able to outperform all individual models in terms of bias, which is the error due to the difference in the average model predictions and actual values, and variance, which is the variability of model predictions, has been the main goal of the studies in this area. An optimization model has been proposed in this paper to design ensembles that try to minimize bias and variance of predictions. Focusing on service sciences, two well-known housing datasets have been selected as case studies: Boston housing and Ames housing. The results demonstrate that our designed ensembles can be very competitive in predicting the house prices in both Boston and Ames datasets.

Keywords: *Machine Learning, Optimal Ensemble, Bias-Variance Trade off, House Price Prediction*

1. INTRODUCTION

The world's economies have been shifted towards service sector in the recent decades. The service related economy accounts for 65% of the world's GDP as of 2017, which has observed a rise from 61% in 2006. In addition, services sector is the leading sector in 201 countries and service related economy consist of more than 80% of total GDP for 30 countries [1]. This has led to more research in the services sector. The term "Service science, management, and engineering (SSME)" was first used by IBM to describe service science as an interdisciplinary approach to the study, design, and implementation of service systems [2]. Service science is defined as "an abstraction of service systems in the same way that computer science is an abstraction of computer-based information systems" [3]. In general, SSME focuses on system design, industry services, software and hardware implementation of service, and etc. [4].

One of the many disciplines of service sciences focuses on information processing services. These services collect, manipulate, interpret, and transmit the data to create value for the end user. Issues such as representation, infrastructure, and self-service are the most significant problems in these services [3].

Machine learning (ML) has been used as one of the powerful tools to deal with the data. Due to the flexibility, machine learning models have been developed in a variety of application domains, from agriculture, bioinformatics, financial trading, fraud detection and smart city management [5]. Several studies have used machine learning algorithms for housing price predictions. In addition, ML models have been implemented on housing datasets for various types of prediction. C4.5 Decision Tree, RIPPER, Naïve Bayes and AdaBoost ML algorithms have been designed to predict Virginia housing prices [6]. A hybrid of genetic algorithm and support vector machines (G-SVM) was proposed in [7] to forecast China housing prices. In another study, SVM was combined with particle swarm optimization (PSO) to forecast real estate prices [8]. Artificial Neural Networks (ANN) and hedonic regression were compared in predicting housing prices in Turkey using a household budget survey data from 2004 [9]. In an empirical study for residential estate appraisal, it was shown that Random Forests perform better than several ML techniques such as CART, KNN, multiple linear regression, ANN, and Boosted trees [10].

Despite of the prediction accuracy achieved by individual ML models, ensemble learning has been proposed to improve prediction accuracy by aggregating predictions of multiple base learners [11]. The ensemble is typically constructed by weighting (in the case of regression) or voting (in the case of classification) the predictions of base learners. The final resulting ensemble often achieves better predictions in comparison to any of single base learners [12]. For instance, the winners of famous real-world data analysis competitions, such as Netflix Prize and KDD Cup 2013, have chosen ensemble approaches as their prediction strategies [13]. The merits of ensemble learners have

* Corresponding author: Tel.: (515) 294-8638; Fax: (515) 294-3524; E-mail: gphu@iastate.edu

generated increasing interests to incorporate this method in research and practice. The results of a comparative evaluation of three ensemble learning methods including Bagging, Boosting and Stacking for credit scoring show the advantage of ensemble learners over base learners [14]. Three financial datasets were chosen to analyze the performance of ensemble learners for classification problem of bankruptcy prediction and it was demonstrated that ensemble learners outperform the best stand-alone method which was multi-layer perceptron neural network [15].

The objective of this paper is to optimize machine learning predictions of ensemble learners by finding the best weights for constructing ensembles for house price prediction. Two housing datasets for Boston and Ames have been chosen to demonstrate and validate the optimization model. Multiple learners including LASSO regression, Random Forests, Deep Neural Networks, Extreme Gradient Boosting (XGBoost), and Support Vector Machines with three kernels (polynomial, RBF, and sigmoid) have been chosen as base learners for prediction. The predictions made by base learners are used as inputs of proposed optimization model to find the optimal weights. The objective is to minimize the Mean Squared Error (MSE) of the predictions, which account for both bias and variance of the predictions.

The paper is organized as follows. The material and methods are introduced in the second section. Section 3 is dedicated to the results and discussions and the paper concludes in the last section on conclusions.

2. MATERIALS AND METHODS

Although shown in various studies that ensemble learners outperform individual base models, designing the optimal method to combine base models remains a significant problem. In many data science competitions, the winners are the ones who could identify the best way to integrate the merits of different models and achieve superior performance.

It has been shown that the optimal choice of weights aims to achieve minimal prediction error by designing the ensembles for the best bias and variance balance. Every predictive model contains error from bias and variance with the amount of each determined by the interaction between the data and model choice. Bias is defined as a model's understanding of underlying relationship between features and target outputs; whereas, variance is the sensitivity to perturbations in training data [16]. Mathematically, for a given dataset $(X, Y) = \{(\boldsymbol{x}, y): \boldsymbol{x} \in \mathbb{R}^{n\times m}, y \in \mathbb{R}^n\}$, we assume there exists a function $f: \mathbb{R}^{n\times m} \to \mathbb{R}$ with noise ϵ such that $y = f(x_i) + \epsilon$ where $\epsilon \sim N(0,1)$.

Using any myriad of supervised learning techniques, we approximate $f(x)$ with $\hat{f}(x)$ [17]. We define the following:

$$Bias\,[\hat{f}(x)] = E\big[\hat{f}(x)\big] - f(x) \tag{1}$$

and,

$$Var[\hat{f}(x)] = E\big[\hat{f}(x)^2\big] - E\big[\hat{f}(x)\big]^2 \tag{2}$$

Taking the mean squared error as the metric of precision, the objective to be minimized can be defined as:

$$E\left[\left(y - \hat{f}(x)\right)^2\right] = \left(Bias\,[\hat{f}(x)]\right)^2 + Var\big[\hat{f}(x)\big] + Var(\epsilon) \tag{3}$$

The third term in the above formula, irreducible error, is the variance of the noise term in the true underlying function which cannot fundamentally be reduced by any model [16].

Naturally, a model with low bias and low variance is desired but not always producible. One common approach to reduce variance among models is to create a bootstrapped aggregated ensemble. Whereas boosting models is used to reduce bias. Each strategy possesses their strength and weaknesses, and finding the optimal balance between the two remains a challenging problem [18],[19].

Taking both bias and variance into account, mean squared error (MSE) has been chosen as the objective function in the mathematical model for optimizing ensemble weights.

$$Min\ \ MSE\left(w_1\hat{Y}_1 + w_2\hat{Y}_2 + \cdots + w_k\hat{Y}_k\ ,\ Y\right) \tag{4}$$

$$s.t.$$

$$\sum_{j=1}^{k} w_j = 1,$$

$$w_j \geq 0, \quad \forall j = 1, \ldots, k.$$

where w_j is the weights corresponding to base model j ($j = 1, \dots, k$), $\hat{Y}_j$ represents the vector of predictions of base model j, and Y is the vector of actual target values. This optimization problem can be formulated as a quadratic programming problem.

$$Min \ \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} w_j \hat{y}_{ij}\right)^2 \tag{5}$$

$$s.t.$$

$$\sum_{j=1}^{k} w_j = 1,$$

$$w_j \geq 0, \quad \forall j = 1, \dots, k.$$

in which, n is the number of observations, y_i demonstrates actual target values of observation i ($i = 1, \dots, n$), and $\hat{y}_{ij}$ is the prediction of observation i by base model j.

This model is a nonlinear convex program. Since the constraints are linear, the convexity can easily be proved by computing the Hessian matrix of the objective function. Therefore, since a local minimum of a convex function on a convex feasible region is guaranteed to be a global minimum, we can conclude that the optimal solution achieves global optimality [20].

We use Python's SciPy optimization library [21] to solve this problem. This library contains numerous algorithms for constrained and unconstrained optimization. For this study, we apply a Sequential Least Squares Programming (SLSQP) algorithm, a special case of sequential quadratic programming [22]. SLSQP utilizes the Han-Powell quasi-Newton method with a BFGS update resulting in robust results to an optimal solution [23].

Three measures have been used to evaluate the model performance. First, mean squared error (MSE) that is a measure of difference between predicted and observed values; second, mean absolute percentage error (MAPE) which expresses accuracy as percentage, and third, the coefficient of Determination (R2) that is defined as the proportion of the variance in the response variable that is explained by independent variables. The R2 ranges from 0 to 1, where values near 1 indicates a perfect fit of predicted values to the observed data.

Validating the results has been done with 10-fold cross validation to estimate the true prediction error. In addition, the hyperparameters in each of models have been tuned by conducting a grid search with 5-fold cross validation.

The proposed optimization model has been applied on two well-known housing datasets. Next two sections describe the details of Boston and Ames housing datasets.

2.1. BOSTON HOUSING

This dataset was collected by the U.S. Census Service regarding housing information in the Boston metropolitan area. The dataset was originally published by Harrison, D. and Rubinfeld, D.L. in a study investigating methodological problems associated with the willingness to pay for clean air using Boston housing dataset [24]. The original dataset is small in size with 506 cases. A description of the variables is presented in Table 1.

Table 1: Boston dataset variables

Variable	Type	Description
CRIM	numeric	Per capita crime rate by town
ZN	numeric	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	numeric	Proportion of non-retail business acres per town.
CHAS	numeric	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	numeric	Nitric oxides concentration (parts per 10 million)
RM	numeric	Average number of rooms per dwelling
AGE	numeric	Proportion of owner-occupied units built prior to 1940
DIS	numeric	Weighted distances to five Boston employment centers
RAD	numeric	Index of accessibility to radial highways
TAX	numeric	Full-value property-tax rate per $10,000
PTRATIO	numeric	Pupil-teacher ratio by town
B	numeric	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	numeric	% lower status of the population
MEDV*	numeric	Median value of owner-occupied homes in $1000's

* target variable

In order to make better predictions, we have scaled the input data to be in the (0,1) range. Then, seven machine learning models including LASSO regression, Random Forests, Neural Networks, XGBoost, and SVM with three kernels (polynomial, RBF, and sigmoid) were applied on the dataset.

2.2. AMES HOUSING

Ames housing dataset was presented by De Cock in 2011 as an alternative to the Boston housing dataset. It describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The original dataset contains 2930 observations and 80 variables. This dataset is used in an ongoing Kaggle data science challenge started in 2016. In this competition, the dataset is split into a train set with size of 1460 observations and a test set of 1459 observations. In this study only the train set of this dataset is used to make predictions. Some of the important variables of this dataset are shown in Table 2.

Table 2: Ames dataset variables (some of the variables are shown here)

Variable	Type	Description
YearBuilt	numeric	Original construction date
Neighborhood	categorical	Physical locations within Ames city limits
Street	categorical	Type of road access
BldgType	categorical	Type of dwelling
MSSubClass	categorical	The building class
Foundation	categorical	Type of foundation
LotArea	numeric	Lot size in square feet
RoofStyle	categorical	Type of roof
Bedroom	numeric	Number of bedrooms above basement level
FullBath	numeric	Full bathrooms above grade
TotalBsmtSF	numeric	Total square feet of basement area
1stFlrSF	numeric	First floor square feet
TotRmsAbvGrd	numeric	Total rooms above grade (does not include bathrooms)
GrLivArea	numeric	Above grade (ground) living area square feet
GarageCars	numeric	Size of garage in car capacity
GarageArea	numeric	Size of garage in square feet
OverallQual	categorical	Overall material and finish quality
ExterQual	categorical	Exterior material quality
KitchenQual	categorical	Kitchen quality
BsmtQual	categorical	Height of the basement
SalePrice*	numeric	The property's sale price in dollars.

* target variable

Pre-processing tasks and data cleanings have been done on this dataset before applying models. These tasks include but not limited to the following:

- Removing outliers observed with two variables (“GrLivArea” and “LotArea”)
- Imputing missing values for many of the variables
- Log-transformation of the target variable
- Log-transformation of the numeric input variables
- Removing highly correlated input variables (“GarageArea”, “1stFlrSF”, and “TotRmsAbvGrd”)
- Constructing three new features with existing variables
- Converting categorical variables to numeric with One-Hot encoding

Afterwards, the following ML models were applied on this dataset to prepare the inputs for proposed optimization:

1. LASSO regression
2. Random Forests
3. Deep Neural Network
4. Extreme Gradient Boosting (XGBoost)
5. Support Vector Machines with polynomial kernel
6. Support Vector Machines with RBF kernel
7. Support Vector Machines with sigmoid kernel

3. RESULTS AND DISCUSSION

The results of each of seven base machine learning algorithms on Boston and Ames housing datasets are presented in Table 3. Each of these models are tuned with a 5-fold cross validation and the error rates shown in the table are estimates of 10-fold cross validation. Based on the results for Boston housing dataset, XGBoost and Random Forests are the best algorithms predicting the median price of the houses with the least MSE and MAPE, and highest R-squared values. In other words, not only these two models predict with highest accuracy, they explain the variation in the target more than other chosen models. Moreover, prediction results of Ames housing dataset finds LASSO and Random Forests as the models with the least MSE and MAPE. These models could explain most of the variations in the target variable with having R-squared values of 0.92 and 0.88, respectively.

Table 3: Base model results for Boston and Ames housing datasets

Error measure	LASSO	Random Forests	Neural Network	XGBoost	SVM (poly)	SVM (RBF)	SVM (sigmoid)
	Boston Housing Dataset						
MSE	35.579	22.479	28.336	21.367	44.315	26.578	34.577
MAPE	20.06%	16.35%	19.99%	16.44%	20.87%	16.06%	18.83%
R^2	0.5785	0.7337	0.6643	0.7469	0.4751	0.6852	0.5904
	Ames Housing Dataset						
MSE	0.0132	0.0183	0.4549	0.0368	0.0275	0.0681	0.0196
MAPE	0.66%	0.77%	3.72%	1.16%	1.00%	1.39%	0.82%
R^2	0.9167	0.8842	-1.8796	0.7669	0.8258	0.5692	0.8758

The prediction vectors of each of the above models are used in the optimization model to find the optimal weight of constructing ensembles with the base learners. Table 4 shows the obtained optimal weights. As it can be seen from the weights in the table below, the optimization model assigns the weight of zero to some models which means that the ensemble excluding these base learners will perform better. Furthermore, the objective function which is the mean square error of the ensemble is less than the MSE of all base learners for both datasets, that shows strength of the optimal ensemble in predicting the targets.

Table 4: Optimal ensemble weights

	w_{LASSO}[1]	w_{RF}[2]	w_{NN}[3]	w_{XGB}[4]	$w_{SVM\text{-}p}$[5]	$w_{SVM\text{-}r}$[6]	$w_{SVM\text{-}s}$[7]	obj[8]
Boston	0	0.113	0.280	0.508	0.073	0.003	0.023	18.901
Ames	0.742	0.221	0	0	0.037	0	0	0.0126

The ensemble with equal weights ($w_i = 1/7$), which is a common practice among data scientists in order to construct ensembles out of some base learners, is considered as a benchmark. The error measures for the ensembles with optimal weights are calculated and compared with the benchmark in Table 5.

[1] LASSO optimal weight
[2] Random forests optimal weight
[3] Neural network optimal weight
[4] XGBoost optimal weight
[5] SVM (polynomial) optimal weight
[6] SVM (RBF) optimal weight
[7] SVM (sigmoid) optimal weight
[8] Objective function value

Table 5: Error rates for optimal and benchmark ensembles

Error measure	Optimal Ensemble	Ensemble with equal weights (1/7)
	Boston Housing Dataset	
MSE	18.901	21.95
MAPE	%15.25	%15.26
R^2	0.7761	0.7399
	Ames Housing Dataset	
MSE	0.0126	0.0281
MAPE	%3.49	%3.34
R^2	0.9199	0.8222

The ensembles with optimal weights outperform the benchmark ensemble as well as each of the base models for both datasets. This is demonstrated in Figure 1. Comparing the error measures, the optimal ensemble has lower MSE, lower MAPE, and higher R^2 value.

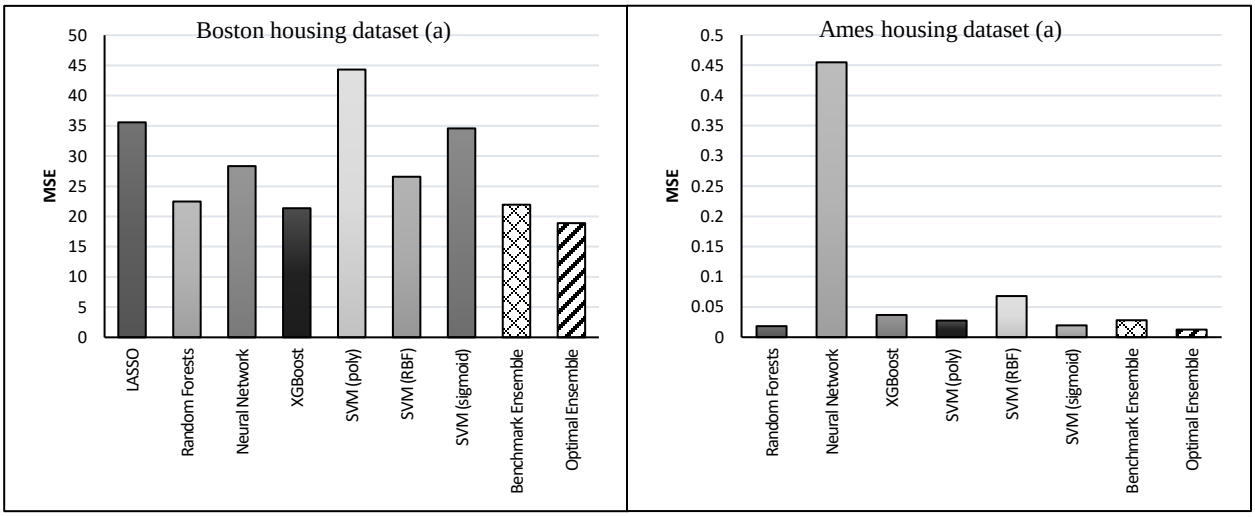

Figure 1: Comparing optimal ensembles with benchmark ensembles and base learners

4. CONCLUSION

A new optimization framework was been proposed in this study which optimizes the mean squared error of multiple base learners to find the optimal weights in designing a new ensemble from them. The designed formulation can result in ensembles that minimize bias and variance of predictions. To validate the performance of the proposed methodology, two famous housing datasets, Boston and Ames datasets, were used as case studies. Seven machine learning algorithms including LASSO, random forests, neural networks and XGBoost along with support vector machines with three kernels were considered as base learners. The created ensembles from the optimal weights found from our optimization model were compared to ensembles created from assigning equal weights to each individual learner and each of the base learners. The results showed that the designed ensemble can outperform the benchmark ensemble as well as all the individual base learners.

The proposed methodology presented a systematic way to find the optimal weights of aggregating predictive learners to create better performing ensembles. This method performed better than each predictive learner in both housing data sets considered in this study. This methodology is generalizable to other data sets in other fields, given that the individual learners are accurate and diverse enough to effectively capture the structure of the data. This diversity in models is the reason of superiority of ensembles. Specifically, having different types of learners (e.g. linear and nonlinear learners) and aggregating these diverse models in a systematic way provides a way to represent different aspects of the data. Hence, this methodology is expected to be generalizable with the ability to predict better compared to initial single base learners.

For the future work, designing a methodology which incorporates finding best ensemble weights while tuning the hyperparameters of each base learner is recommended. This method can find the best hyperparameters and optimal weights of creating ensemble at the same time.

REFERENCES

[1] *World development indicators. World Bank, 1978.*

[2] *B. Hefley and W. Murphy, Service science, management and engineering: education for the 21st century. Springer Science & Business Media, 2008.*

[3] *H. Katzan, "Foundations of service science concepts and facilities," Journal of Service Science, vol. 1, no. 1, 2008.*

[4] *G. Xiong, Z. Liu, X. Liu, F. Zhu, and D. Shen, Service Science, Management, and Engineering:: Theory and Applications. Academic Press, 2012.*

[5] *L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," Statistical science, vol. 16, no. 3, pp. 199-231, 2001.*

[6] *B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," Expert Systems with Applications, vol. 42, no. 6, pp. 2928-2934, 2015.*

[7] *J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," Expert Systems with Applications, vol. 38, no. 4, pp. 3383-3386, 2011.*

[8] *X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," Optik-International Journal for Light and Electron Optics, vol. 125, no. 3, pp. 1439-1443, 2014.*

[9] *H. Selim, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," Expert systems with Applications, vol. 36, no. 2, pp. 2843-2852, 2009.*

[10] *E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics," Expert Systems with Applications, vol. 39, no. 2, pp. 1772-1778, 2012.*

[11] *P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining, Addison," ed: Boston, MA USA: Wesley Longman, Publishing Co., Inc, 2005.*

[12] *D. Talia, P. Trunfio, and F. Marozzo, Data analysis in the cloud: models, techniques and applications. Elsevier, 2015.*

[13] *M. Sugiyama, Introduction to statistical machine learning. Morgan Kaufmann, 2015.*

[14] *G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," Expert systems with applications, vol. 38, no. 1, pp. 223-230, 2011.*

[15] *L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," Expert systems with applications, vol. 36, no. 2, pp. 3028-3033, 2009.*

[16] *J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning (no. 10). Springer series in statistics New York, 2001.*

[17] *L. Breiman, "Bias, variance, and arcing classifiers," 1996.*

[18] *H. Pham and S. Olafsson, "Bagged ensembles with tunable parameters," Computational Intelligence, vol. 35, no. 1, pp. 184-203, 2019.*

[19] *H. Pham and S. Olafsson, "On Cesaro Averages for Weighted Trees in the Random Forest," Journal of Classification, 2019.*

[20] *S. Boyd and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.*

[21] *E. Jones, T. Oliphant, and P. Peterson, "others. SciPy: Open source scientific tools for Python," W eb http://www.scipy.org, 2001.*

[22] *D. Kraft, "A software package for sequential quadratic programming," Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt, 1988.*

[23] *A. Wendorff, E. Botero, and J. J. Alonso, "Comparing Different Off-the-Shelf Optimizers' Performance in Conceptual Aircraft Design," in 17th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 2016, p. 3362.*

[24] *D. Harrison Jr and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," Journal of environmental economics and management, vol. 5, no. 1, pp. 81-102, 1978.*