2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress

Multiple Linear Regression with Kalman Filter for Predicting End Prices of Online Auctions

Xiaohui Li
College of Computer Science and Technology
Harbin Engineering University
Harbin, China
lxhhrb@hrbeu.edu.cn

Hongbin Dong*
College of Computer Science and Technology
Harbin Engineering University
Harbin, China
donghongbin@hrbeu.edu.cn

Shuang Han
College of Computer Science and Technology
Harbin Engineering University
Harbin, China

***Abstract*—In the whole online auction industry, it is important to predict end prices. To improve the accuracy of predicting end prices using less training data, this paper proposes a hybrid algorithm which combines multiple linear regression with Kalman filter (MLRKF). The proposed algorithm solves problems of low prediction accuracy and over fitting when we signally use multiple linear regression algorithm to predict in machine learning. Firstly, multiple linear regression and Kalman filter models are introduced and analyzed. Secondly, we view the prediction problem with multiple linear regression as a weight parameter optimization problem, and demonstrate the method in theory. Then MLRKF prediction model is provided. Finally, the proposed model is used to predict eBay end prices based on two datasets. In our experiment, MLRKF prediction model has been compared with other models including multiple linear regression, multiple linear ridge regression, Lasso, random forest, support vector machine, and recurrent neural network. This hybrid algorithm has been proved to produce highly accurate results with less training data and lower time cost. The experimental results indicate that the proposed algorithm has a small error rate by calibration metrics in behavioral bidding tasks.**

Keywords-multiple linear regression with Kalman filter; predicting end prices; online auctions

I. INTRODUCTION

With the prosperity of global e-commerce, there are many opportunities and challenges for new models and new businesses. Online auctions are an important part of e-commerce activities. Many firms can be offered a great benefit by predicting end prices of online auctions in both business-to-business and business-to-consumer markets [1], so machine learning algorithms are used to predict end prices. eBay is a leading online auction and shopping website in the world. All the people around the world can buy and sell goods on eBay website. On each day, eBay offers millions of items for sale and produces a large amount of transaction data during commodity transaction. The huge auction data from eBay website can be exploited to provide service for sellers and buyers. Machine learning technology and several data mining methods have been used in the area of forecasting auction end prices [2-6].

Predicting the end price of an online auction is an important and challenging research area, because it has an impact on bidding participants' revenue by using predicted prices. With detecting auction fraud, accurate price prediction could give suggestions to reduce money loss of all honest bidders in auctions [7-9]. Therefore, participants expect accurate prediction results in an online auction market.

Machine learning technology has been applied to many prediction methods and social network service [10-12]. Khadge [13] proposed a system, which could collect huge auction data from eBay and predict end prices of online auctions using machine learning algorithms. The proposed system used Naïve Bayes and support vector machine to predict classification accuracy and maximize profit or not respectively. The algorithm accuracy was 99.33% in classification of whether the item will sell or not and 96.3% accuracy for predicting whether an item maximize profit or not. Gupta [14] proposed a machine learning framework for predicting purchase by online customers based on dynamic pricing. The results of the proposed model were compared with other techniques. In order to solve prediction problems, logistic regression [15], Bayesian linear regression [16], decision trees [17] and deep recurrent neural network [18] have been also used. Multiple linear regression analysis is a simple method to solve regression problems, which can accurately measure the correlation degree and regression fitting degree of factors in a system.

A simple linear regression model may be inaccurate and the prediction errors are almost unacceptable. Combining linear regression algorithm and artificial neural network can provide good results in prediction problems, but there are inherent over fitting characteristics. Therefore, they may not be the best way to predict time series and development trend.

Kalman filter is a key tool for time series prediction and analysis [19-21]. Kalman filter method for predicting has an advantage of dynamically modifying prediction weights, and it can obtain higher accuracy by relying on prediction recurrence equations. Therefore, Kalman filter can be

978-1-7281-6609-4/20/$31.00 ©2020 IEEE
DOI 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00042

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

approximated by a regression of recent observed values. Kalman filter algorithm is widely used in control fields. It is adopted to estimate unmeasurable variables and remove the noise in measurements. Chen [22] predicted the unmanned aerial vehicle information using unscented Kalman filter algorithm. An auction algorithm was used to clear up the collision. After that, they assigned the tasks. The results showed that the proposed algorithm was better than the greedy task assignment and consensus algorithm.

The observed data contains some noise and interference from an online auction platform, so predicting prices can also be regarded as a filtering process. In order to use less training data and improve prediction accuracy, this paper proposes a multiple linear regression with Kalman filter (MLRKF) prediction model. Due to the addition of filtering, the model can solve the over fitting problem of multiple linear regression algorithm in machine learning. When new auction data is obtained, the weight coefficient state variables can be updated recursively in our model using MLRKF algorithm. This hybrid algorithm can modify the linear regression model to a certain extent and improve its prediction accuracy.

The idea motivating this paper is that eBay website data extraction along with a necessarily simple but sufficiently accurate MLRKF model can be used to predict auction end prices. In the process of integrating prediction and updating stages of Kalman filter into a multiple linear regression algorithm, regression weight coefficients are updated by the established Kalman filter model, and corresponding parameters of Kalman filter model are adjusted dynamically to make the model accurately predict auction end prices. In this paper, the application of MLRKF prediction model is presented in eBay online auctions. However, the proposal is intended to be applicable to more study of forecasting prices in any online auction market.

At present, it is necessary and practical to combine different models and investigate different approaches. According to the aforementioned problems, the specific contributions are shown as follows.

- In order to develop the advantages of different prediction models with higher accuracy and smaller error, MLRKF prediction model is proposed, which can dynamically optimize regression weight coefficients. This hybrid algorithm can solve the problems of large demand for training data, large time cost and over fitting.
- Compared with other hybrid intelligent algorithms, MLRKF is simple to model and can obtain the model explicit expression. Kalman filter is rarely used in the field of predicting online auction end prices. MLRKF can help to make up for the deficiency of a single prediction algorithm in online auctions.
- MLRKF prediction model is used to predict eBay auction end prices, and compared with various models including multiple linear regression, multiple linear ridge regression, Lasso, random forest, support vector machine, and recurrent neural network. The experimental results demonstrate that MLRKF prediction model can produce highly accurate results with less training data and lower time cost.

The rest of the paper is introduced as follows: Section II introduces the related work. Section III shows the basis of extrapolative model in theory. The general form of parameter optimization problem is introduced. Section IV introduces MLRKF prediction model and algorithm, which are used to predict auction end prices, and then improves the model. Section V introduces calibration metrics. We analyze the experimental results and discuss the detailed results in Section VI. Finally, conclusions and new future work are discussed in Section VII.

II. Related Work

There are many applications in the prediction field of multiple linear regression. Li [23] used BP network, multivariate regression and logistic regression methods to forecast the final prices of auction items. The experimental results showed that neural network was better than logistic regression traditional statistical method in handling highly skew data. Díaz [24] proposed a regression tree method for modeling electricity price formation. The model showed good accuracy in predicting the price formation. Wang [25] presented a multiple criteria linear programming regression (MCLPR) prediction model for predicting click-through rate. The model was compared with support vector regression and logistic regression in the experiment. The results demonstrated that the proposed model MCLPR was an efficient method in behavioral targeting tasks. Li [26] developed a novel optimization framework for learning to predict prices using the shrinkage method, named as E-commerce online auction machine. Zhang [27] used the hedonic regression approach to select key variables and found the relationships between key variables and the final winning price. Linear regression models [28-30] have been widely used in trending analysis. For example, they have been proven to be good fits for estimating the growth of natural resources [31] and predicting population growth [32,33]. A conventional linear regression model was applied for trend analysis with slope coefficients in [34,35]. It is one of the most challenging problems to accurately define a multivariate linear model to find the relationship between dependent variables and independent variables.

There are also many achievements in the applications of linear regression and other methods. Albert [36] provided a new idea of combining local optima based linear structures. A multilinear weighted regression with neural networks was proposed for trend prediction. Arsić [37,38] used multiple linear regression and the artificial neural networks methods to predict ozone concentration in ambient air. Compared to the multilinear regression model, the results showed that ANNs provided better estimates of ozone concentration on the monitoring site. Wu [39] studied the ad inventory allocation and predicted advertisers' bidding prices. The authors proposed a hybrid model, which combined linear regression on bids with observable winning prices and censored regression on bids with the censored winning prices. The proposed model was weighted by the winning rate of the DSP.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

Kalman filter is also widely used in the prediction field. Matzuka [40] used Kalman filtering to predict time-varying parameters in a model. Liu [41] proposed a rapid algorithm for quickly discovering neighbor nodes in dynamic environment. The algorithm was based on a novel mobility prediction model using Kalman filter theory. Each vehicular node had a prediction model to predict its own and its neighbors' mobility in the discovery algorithm. Huang [42] proposed a hybrid model to improve the train running time prediction accuracy during railway disruptions. The hybrid model comprised support vector regression and Kalman filter in machine learning models.

Machine learning methods focus on how some variables can be used to predict others. A linear regression obviously focuses on determining the impact of some variables on others. Whereas artificial neural networks, support vector machines and random forests are less sensitive to the problems shown by traditional linear regression techniques in machine learning techniques [24]. When we predict that whether an auction will be successful or not, this problem can be seen as a binary classification. Logistic regression is often used to predict binary classification problems. When the models are trained with less training data, Bayesian linear regression has greater uncertainty, and decision trees tend to over fit. A lot of data is needed to train deep recurrent neural network models. The models are in the black box state, so it is difficult to understand the internal mechanism. We need fully analyze the characteristics that affect the target value. Multiple linear regression is direct and fast, but it is easy to cause over fitting. In order to solve this problem, we combine it with Kalman filter.

III. METHODS

A. Multiple Linear Regression

Multiple linear regression analysis can improve the effect of a prediction equation, so it is more suitable for practical economic problems. In real economic problems, few variables are affected by a single factor, but more by combinations of multiple factors. Therefore, we can use multiple linear regression analysis method to solve the online auction analysis and prediction problems.

We find that the change of online auction end prices is affected by several factors. There is a relationship between one dependent variable and some other independent variables. Regression analysis offers an effective relating express, that is the response of a dependent variable to a set of observation independent variables. Multiple linear regression is a universal model in mathematical statistics, and it is also a mathematical method to deal with multivariate relations. In economic problems, a variable is often affected by multiple variables, so the most direct way to conduct such analysis is the means of multiple linear regression model. When the relation between independent variables and a dependent variable is linear, the general form of multiple linear regression model can be expressed in (1):

$$y(x) = y_i(x) = X_i\beta + \varepsilon_i \quad (1)$$

where y_i is a predicted value, $X_i = (1, x_1, x_2, ..., x_i)$ is a vector of descriptive variables, $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ is a vector of coefficients, and ε_i is a random error for different variables.

The dependent variable is calculated by using (1). In our case, y_i is the end price of an online auction *i*. $x_1, x_2, ..., x_i$ $(i = 1, 2, ..., 4)$ are the available explanatory variables, and they may be StartingBid, HitCount, AuctionAvgHitCount, and SellerItemAvg.

In order to solve issues of over fitting and irreversibility in linear regression, multiple linear ridge regression prediction (MLRRP) is presented as Algorithm 1.

Algorithm 1: MLRRP Algorithm

Input: Loading Online Auction Dataset from E-commerce Website
Output: Regression weights *ws* , Estimate Value *yhat*

Training set $X_0 = [x_1, x_2, \ldots, x_n; y]$

Set classifier function

$y(x) = y_w(x) = w_0 + w_1x_1 + ... + w_ix_i \quad (i = 1, 2, ..., n)$

Parameter w :regression weights m: sample number

Initialize:

Initialize matrix: x [], y []

Create a diagonal weight matrix: *weights* []

Error estimating function $\theta(w) = \sum_{i=1}^{n} (y_i - x_i^T w)^2$

Repeat:

standardized data

for j in range (downsize numTestPts)

$ws = ridgeRegress()$

$wMat[i,:] = ws^T$

Calculate $\theta(ws) = (X^T X + \lambda I)^{-1} * X^T y$

If $X^T X$!=0 get *ws*

End

B. Kalman Filter

Kalman filter is a linear minimum variance estimation algorithm for a state sequence in a dynamic system at first. In a dynamic observation system, it can accurately predict the position and speed of the target through the optimal estimation based on the input and output data. Therefore, Kalman filter plays an important role in estimating a past value (a interpolation or smoothing process), a current value (a filtering process) and a future value (a predicting process).

Liang [43] proposed a fusion method combining Kalman filtering and K-nearest neighbor approach. In experiment, they used the fusion model to predict short-term passenger flow prediction in urban public. Chen [44] proposed a novel hybrid algorithm based on interference suppression and Kalman filter. The proposed algorithm can achieve higher accurate and robust in real-time tracking than Kalman filter only.

Kalman filter theory is divided into three contents: a filtering problem, a prediction problem and a smoothing problem. This paper focuses on the prediction problem. In order to apply Kalman filter to prediction, a recurrence equation of prediction must be derived. It is usually derived by using orthogonal theorem and mathematical induction. Then a prediction recurrence equation is obtained.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

Kalman filter algorithm is divided into two steps: prediction and update.

Step 1 (Prediction). The state of the current time (time k) is estimated according to the posterior estimation of the previous time (time $k-1$), and the prior estimation of time k is obtained.

Step 2 (Update). Use the measured value at the current time (time k) to correct the estimated value at the prediction stage and get the posterior estimated value at the current time (time k).

Kalman filter can be divided into a time renewal equation and a measurement renewal equation, which are also called a prediction equation and a correction equation. So Kalman filter algorithm is a recursive predicting-correcting method.

The core of Kalman filter includes two processes and five formulas.

1) Prediction process is defined as follows:

$$\hat{x}_k = F_k \hat{x}_{k-1} + B_k \vec{u}_k \tag{2}$$

$$P_k = F_k P_{k-1} F_k^T + Q_k \tag{3}$$

2) Update process is defined as follows:

$$\hat{x}_k^{'} = \hat{x}_k + K^{'}(\vec{z}_k - H_k \hat{x}_k) \tag{4}$$

$$P_k^{'} = P_k - K^{'} H_k P_k \tag{5}$$

$$K^{'} = P_k H_k^T (H_k P_k H_k^T + R_k)^{-1} \tag{6}$$

where $\hat{x}_{k-1}$ and $\hat{x}_k^{'}$ are updated results of filtering at time $k-1$ and k. $\hat{x}_k$ is the prior state estimation at time k, and $\hat{x}_{k-1}$ is predicted according to the optimal estimation at previous time $k-1$. $\hat{x}_k$ is the result of prediction equation. P_{k-1} and $P_k^{'}$ are posterior estimation covariance at time $k-1$ and k respectively, that represent the uncertainty state. P_k is prior estimation covariance at time k, which is an intermediate result of filtering.

H_k is a transformation matrix from some state values to measurement variables. In Kalman filter, it is a linear relationship. It is responsible for transforming *m* dimension measurement values to *n* dimension values, that is one of filtering preconditions. $\vec{z}_k$ are measured values (a vector of observed values), which are input data of a filter. $K^{'}$ is a filtering gain matrix, which is an intermediate calculation result. It is also called Kalman gain or Kalman coefficient. F_k is a transition matrix to transfer state vector from one state to other, which is a conjecture model for the target state transition. Q_k is process excitation noise covariance (covariance of system process), which is used to represent errors between a state transition matrix and an actual process. We cannot observe process signal directly, so it is difficult to determine Q_k value. Q_k is used to estimate state variables of discrete-time process, which is also known as noise caused by a prediction model. It is a covariance matrix of state transition.

R_k is a measurement noise covariance matrix. When the filter is implemented, R_k can be generally observed. It is a filter known condition. B_k is a matrix that converts input data to a state matrix.

C. Optimization on Parameters

Kalman filter is used to optimize weight coefficients, which is a parameter optimization problem. Let X_T denote the training dataset, and X_V denote the test dataset. Assuming that datasets X_T and X_V obey the natural distribution G_X, we treat the given problem F as a learning problem, which is composed of one or more different steps or algorithms. For a given online auction, let $\theta=(w_1, w_2, ..., w_n)$ denote parameters of problem F, where n is the number of parameters, and w_i is a feature weight. The parameter w_i represents a parameter to be optimized. For given X_T and θ, model f represents a function expression $f = F(\theta, X_T)$, which is obtained by solving problem F. In the optimization process, quantitative evaluation criterion $g(\bullet)$ is used to measure the model *f*. For predicting end prices, the mean absolute percentage error is mainly used. The formal expression of feature weight optimization is as follows:

$$\theta_{opt} \approx \arg_{\theta} \min \frac{1}{n} \sum_{x \in X_V} g\left(f_{\theta, X_T}(x)\right) \tag{7}$$

In parameter optimization process, parameter sampling and verification are the considered cost of calculation. In the limited parameter sampling condition, training model takes up the main time cost. Finding optimal feature weights are the main goal of our feature weight optimization. Kalman filter can deal with the state estimation of multivariable dynamic nonlinear systems, and is suitable for machine learning parameter optimization [45].

D. Expounding and Proving method

Using Kalman filter to deal with parameter estimation can be expressed as: given a dynamic system and a group of system observed data, we need to find a joint probability density function to describe parameters or system state.

Let x denote the *n*-dimensional weight vector of a multiple linear regression algorithm. We assume that x follows Gaussian distribution with mean x_b and covariance P_b. x_b is a prior estimate of x. Observed z represents the expected output of a trained learning model on a given test dataset. The observed operator H is used to associate feature weights with a prediction result on a test dataset, and its value corresponds to observed data. It is assumed that the observation obeys Gaussian probability distribution function with mean Hx and covariance R. According to Bayesian theory, the posterior probability density of parameter vector x can be defined as $P(x) \propto \exp(-F(x))$, where $F(x)$ is defined as follows:

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

$$F(x)=\frac{1}{2}\left((x-x_b)^T P_b^{-1}(x-x_b)+(z-Hx)^T R^{-1}(z-Hx)\right) \quad (8)$$

By minimizing *F(x)* in (8), a feature weight vector x_a can be found to maximize the posterior probability density. x_a and *K* can be defined as follows:

$$x_a = x_b + K(z - Hx_b) \quad (9)$$

$$K = P_b\hat{H}^T\left(\hat{H}P_b\hat{H}^T + R\right)^{-1} \quad (10)$$

where $\hat{H}$ is tangent linear approximation of observation *H*, and the covariance of an analysis parameter vector x_a is expressed as follows:

$$P_a = \left(I - K\hat{H}\right)P_b \quad (11)$$

According to (10), when observation covariance *R* is small, the covariance of a parameter vector will decrease rapidly in each analysis step. It shows that the calculation is converging rapidly.

IV. MLRKF Prediction Model and Algorithm

The MLRKF prediction workflow can be showed in Fig.1.The workflow of MLRKF prediction algorithm can be described as following steps:

Step 1. In training data, weight coefficients of related variables are obtained by a multiple linear regression algorithm.

Step 2. Filter, initial state vector and other parameters are Initialized.

Step 3. The weight coefficients corresponding to a certain percentage of training data are regarded as observation values, and the weight coefficients corresponding to all training data are regarded as the measurement value. The iterative calculation starts until the convergence, and output values are weight coefficients obtained.

Step 4. We can use the output values of weight coefficients to predict test data, count prediction errors, and measure prediction effect.

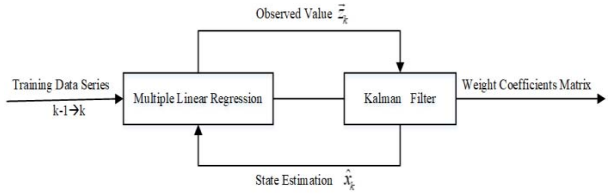


Figure 1. Flow graph of MLRKF prediction algorithm.

A novel optimization framework is proposed, named as MLRKF prediction framework as shown in Fig.2. In this framework, the proposed algorithm can be used to calculate regression weights for predicting auction end prices. The basic idea of MLRKF is: Firstly, an online auction system framework is introduced and modeled, then the experimental data from eBay website is divided into training dataset and test dataset according to a certain proportion. Secondly, the corresponding regression weight coefficients are obtained by multiple linear regression algorithm. Finally, a certain proportion of the training dataset (or the whole training data) and the whole training dataset (or the whole dataset) are regarded as observation values and measurement values respectively. Then the regression weight coefficients are finally obtained by the recursive program of Kalman filter prediction. Experimental results demonstrate that the hybrid algorithm can predict auction end prices more accurately using less training data. The prediction accuracy is higher than that of other multiple linear regression models. MLRRP algorithm, which is a simple improvement of multiple linear regression prediction (MLRP) algorithm, is shown as Algorithm 1. For ease of reference, we summarize MLRKF prediction algorithm as Algorithm 2.

MLRKF prediction model should be presented as follows:

$$y(x) = y_w(x) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_ix_i (i = 1,2,\ldots,n) \quad (12)$$

$$w_i = F_K w_i' \quad (13)$$

$$\hat{w}_i = w_i + K'\left(\bar{z}_k - H_k w_i\right) \quad (14)$$

where w_0 is a constant. $w_i (i = 1,2,\ldots,n)$ is a regression coefficient. In the filtering process, $w_i (i = 1,2,\ldots,n)$ is a prior state estimation at bidding time and an intermediate calculation filter result. w_i' is one of filtering results, that is, the updated result, also known as optimal estimate.

According to the optimal estimation at training time, w_i is a predicted result at test time and is also a result of a prediction equation. w_i' is a posterior state estimate at training time, which is one of filtering results. $\hat{w}_i$ is a posterior state estimate at testing time, which is one of filtering results. F_k is a state transition matrix. K' is a filter gain matrix, which is an intermediate calculation filtering result, also known as Kalman gain. $\bar{z}_k$ is a training value and it is also an input filtering value. H_k is a transformation matrix from a state variable to a measurement variable. In Kalman filter, it is a linear relationship. It is responsible for transforming *m* dimension measurement values to *n* dimension values, that is one of filtering preconditions. $\bar{z}_k - H_k w_i$ are residual of actual observation values and prediction observation values, which can be used to modify the prior (prediction) to obtain the posterior with the Kalman gain.

V. Calibration Metrics

For numerical prediction problems, researchers often care about prediction accuracy. In current study, the criteria used to measure forecasting accuracy are mean absolute error (MAE), root mean-square error (RMSE), and mean absolute percentage error (MAPE).

MAE is defined as follow:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|E_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (15)$$

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

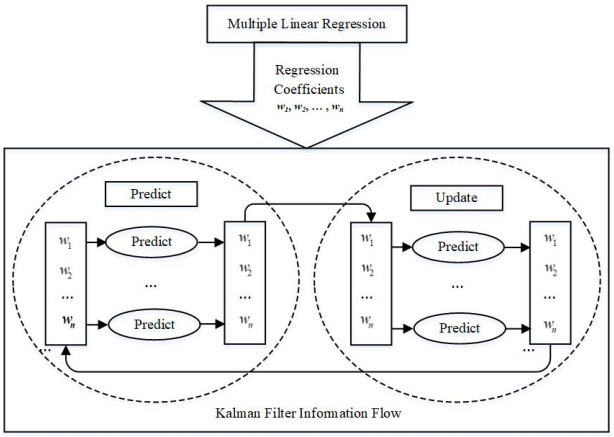


Figure 2. The framework of MLRKF prediction.

Algorithm 2: MLRKFP Algorithm

Input: Loading Online Auction Dataset from E-commerce Website
Output: Regression weights state_pre[w], Estimate Value *yhat*

Training set X_0=[x_1,x_2,...x_n;y]

Set classifier function $y(x) = y_w(x) = w_0 + w_1x_1 + \ldots + w_ix_i$

$(i = 1, 2, \ldots, n)$

Parameter w :regression weights m: sample number

Initialize:

Initialize matrix: x[], y[]

Error estimating function $\theta(w) = \sum_{i=1}^{n} (y_i - x_i^T w)^2$

Repeat:

Calculate $X^T X$

If $X^T X$!=0 get w

for w in range(1, n_iter):

state_pre[w] = state_kalman[w - 1]

Pminus[w] = P[w - 1] + Q

K[w] = Pminus[w] / (Pminus[w] + R)

state_kalman[w]= state_pre[w] + K[w] * (z[w] - state_pre[w])

P[w] = (1 - K[w]) * Pminus[w]

get state_pre[w]

End

RMSE is defined as follow:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n} E_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (16)$$

RMSE is a widely used numerical prediction evaluation index.

MAPE is defined as follow:

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (17)$$

where y_i is the actual value of sample i, $\hat{y}_i$ is the estimate of sample i, and n is the total number of samples.

MAPE is an average of relative errors absolute sum. Compared with simple relative error, it avoids a problem that the positive and negative relative errors cannot be added. At the same time, by calculating an average, it also reflects an average level of prediction relative errors. It is a frequently used numerical prediction evaluation index. If MAPE is less than 10%, we think the prediction effect is very good usually.

VI. Experimental Results and Discussions

A. Data

Two datasets of eBay auctions are used in experiment. Dataset1 is loaded from https://cims.nyu.edu/~munoz/data/. The dataset can be used for evaluating second price auctions with reserve. It contains four data files, which are described in Table I. The main feature names and descriptions of dataset1 are shown in Table II.

TABLE I. Data File Descriptions

Data file name	File description	Data rows
TraingSet	All auctions in April 2013	258588
TestSet	All auctions in the first week of May 2013	37460
TrainingSubset	All auctions successfully traded in April 2013	79732
TestSubset	All auctions successfully traded in the first week of May 2013	9392

Fig.3 shows the correlation between two different characteristics. Through the second column in Fig.3, we can see the correlation between different characteristics and price. Characteristics including StartingBid, AvgPrice, HitCount, AuctionAvgHitCount, AuctionMedianPrice, AuctionCount, AuctionHitCountAvgRatio, and SellerItemAvg have a strong positive correlation with price, while category has a large negative correlation with price. Fig.4 shows that there is a positive correlation between AuctionSaleCount and AuctionCount. Price, AuctionSaleCount, AuctionCount and StartingBid show a serious right skew distribution.

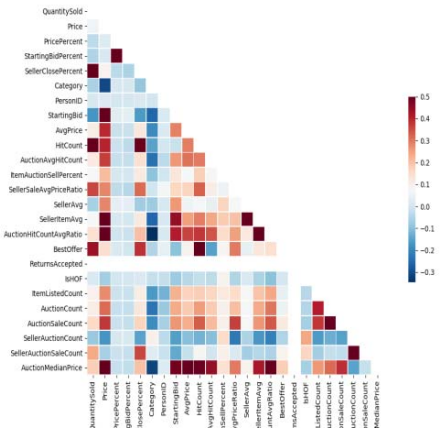


Figure 3. The illustration of impact from bid characteristics on all auctions.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

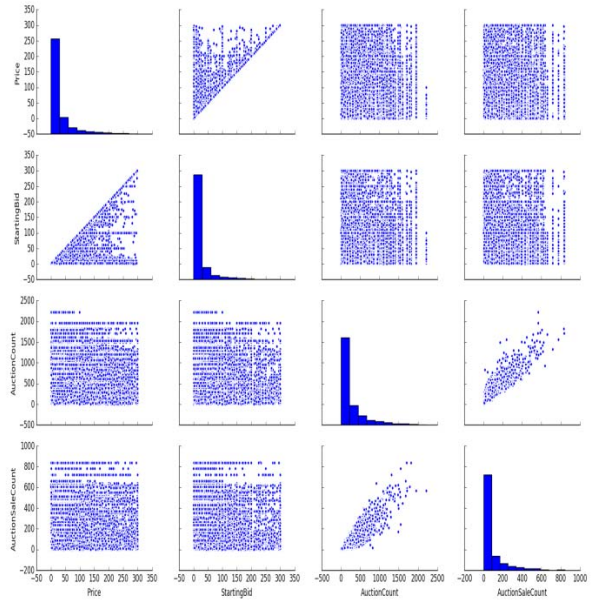


Figure 4. The illustration of impact from bid characteristics.

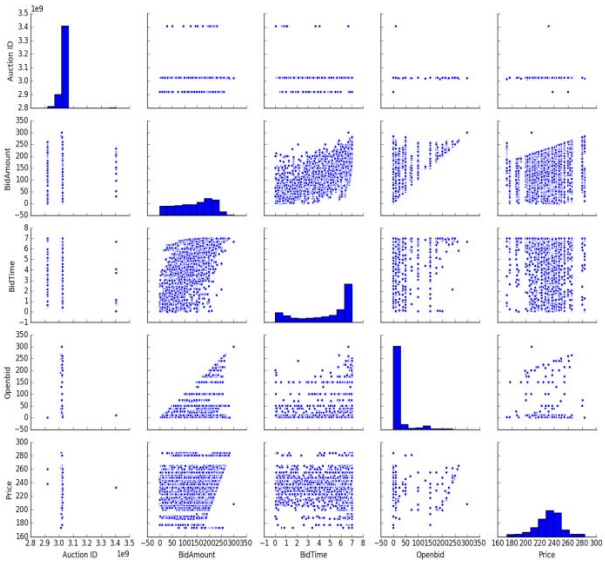


Figure 5. The illustration of impact from bid characteristics on Palm Pilot PDAs auction.

Dataset2 is a real-world sample dataset with eBay auctions on Palm Pilot PDAs. Fig.5 is a scatter matrix of auction characteristics, which illustrates the impact of bid characteristics. The diagonal is the histogram of characteristic variables. Through the histogram, we can see that price follows a positive and negative distribution, and Openid shows a serious right skew distribution. The price histogram illustrates that price obeys normal distribution. There are almost no outliers. The fields used for the research are shown in Table III.

TABLE II. DATA FEATURE DESCRIPTIONS

Feature name	Feature description
Price	end prices of auctions
StartingBid	minimum transaction price of an auction
BidCount	number of bids won in an auction
Title	transaction title
QuantitySold	successful sale number (0 or 1)
SellerRating	seller's rating on eBay
StartDate	auction start date
EndDate	auction end date
PositiveFeedbackPercent	percentage of positive feedback received by seller (for all feedback)
BuyitNowPrice	price for immediate purchase
HighBidderFeedbackRating	eBay rating of the highest-price bidder
IsHOF	the seller is or not a hall of fame player (0 or 1)
AvgPrice	average price of a good in inventory
MedianPrice	median price of a good in inventory
AuctionCount	total number of auctions in inventory
SellerSaleToAveragePriceRatio	proportion of auction goods price to average price
AuctionDuration	auction duration days
StartingBidPercent	the ratio of the starting bidding price to the average transaction price
ItemAuctionSellPercent	percentage of successful auctions in all online auctions

TABLE III. THE FIELDS AND DESCRIPTORS ON PALM PILOT PDAS

Field	File description
Auction ID	unique identifier of an auction
Bid Amount	the proxy bid placed by a bidder
Bid Time	the time (in days) that the bid was placed, from the start of the auction
Bidder	eBay username of the bidder
Open bid	the opening bid set by the seller
Price	the closing price that the item sold for (equivalent to the second highest bid + an increment)

B. Results

- Dataset 1

As the most extensively studied online market, eBay provides a large amount of transaction data resource for the study of auctions in general. MLRKF is an improvement of MLRP and MLRRP algorithms, so we only draw the graph of these algorithms.

The proposed model is also compared with models including random forest regressor, SVM, RNN and Lasso that have been previously applied for forecasting auction end prices. The training time of each model with dataset1 is shown in Fig.6. We can see that MLRKF has the lowest training time cost. Table IV shows that the proposed model has the smallest MAE, RMSE, and MAPE values compared to MLRP, MLRRP, Lasso, Random Forest, SVM and RNN

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

models. As tabulated in Table IV, the results demonstrate that the proposed model has the highest accuracy validated using the three metrics.

We can see that the error rate of MLRP algorithm is the highest. The error rate of MLRRP algorithm is relatively higher than that of Lasso, random forest, SVM and RNN. However, after Kalman filtering, the error rate is greatly reduced. We use a simple and visible optimization model to get better prediction results.

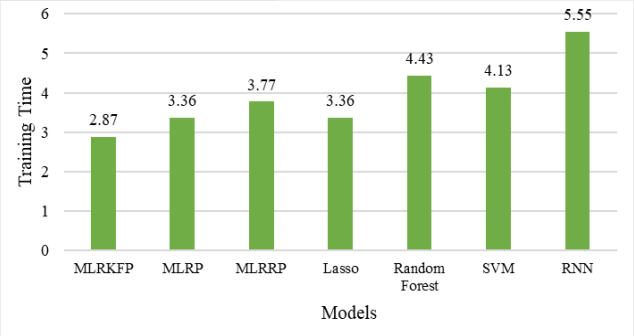


Figure 6. Training time of each model using dataset1.

TABLE IV. ERROR COMPARISON OF DIFFERENT MODELS

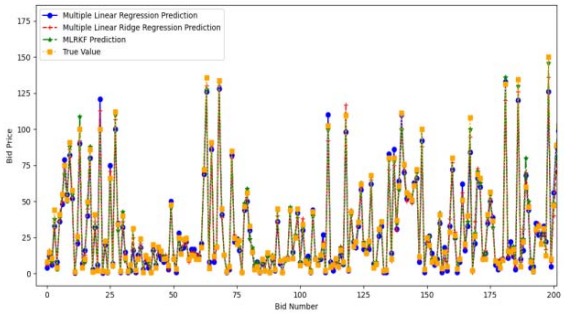
Models	MAE	RMSE	MAPE
MLRKFP	2.63	3.56	5.28
MLRP	5.42	5.80	5.43
MLRRP	5.37	5.74	5.40
Lasso	3.25	4.50	6.99
Random Forest	3.21	4.55	6.52
SVM	3.22	4.42	6.54
RNN	3.21	4.38	6.37

In order to display predicting results clearly, 800 predicting results are randomly selected to plot. Prediction values of MLRP, MLRRP and MLRKF models vs. true values are plotted as shown in Fig.7. We can find MLRKF prediction plot is better fitted for true end prices with dataset1.

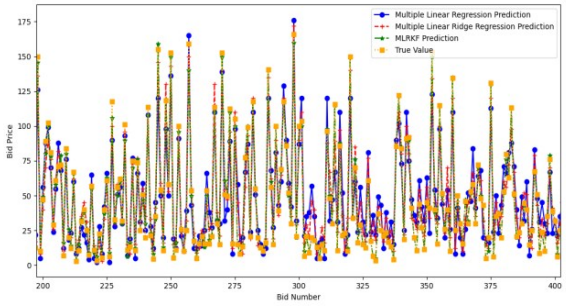
- Dataset 2

There are only 3646 rows data in dataset2. In this paper, we used 70% data of eBay auctions on Palm Pilot PDAs for training and 30% for testing. Less training data is used, but prediction results are better. Because Kalman filter prediction method can dynamically modify prediction weight coefficients, we extract 50% of the training data as observation values and 70% of the training data as measurement values in MLRKF prediction algorithm. Although the training data is reduced, we can still get better prediction accuracy.

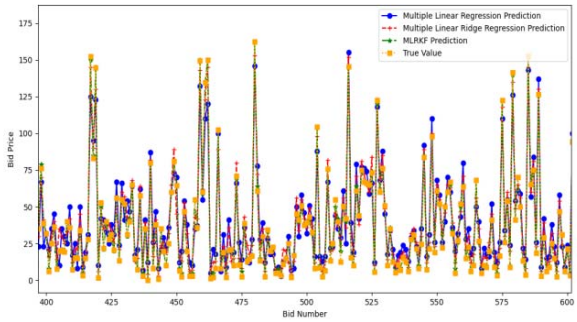
We can find MLRKF prediction plot is better fitted for true bidding end prices with eBay auctions dataset on Palm Pilot PDAs. Fig.8 shows that the proposed model positively follows the stochastic change in bidding end prices while other models fail. The plot also depicts that the proposed model accurately follows the price trends with bid number.



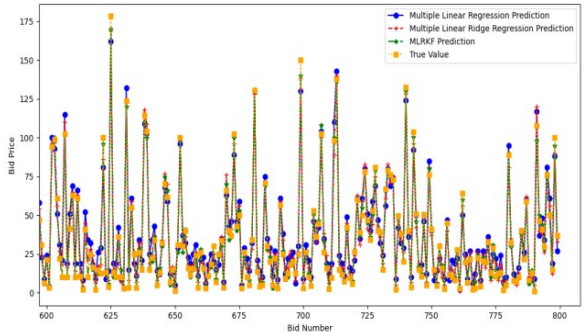
(a) Bid number is between 0 and 200.



(b) Bid number is between 200 and 400.



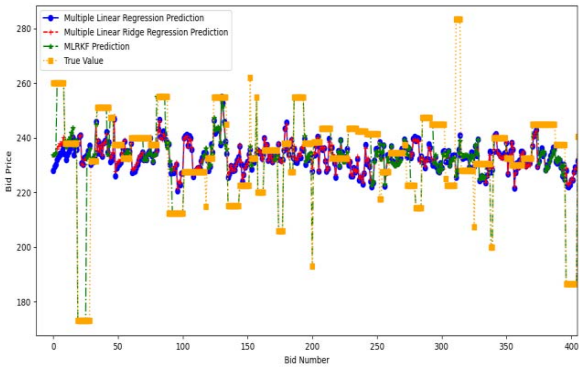
(c) Bid number is between 400 and 600.



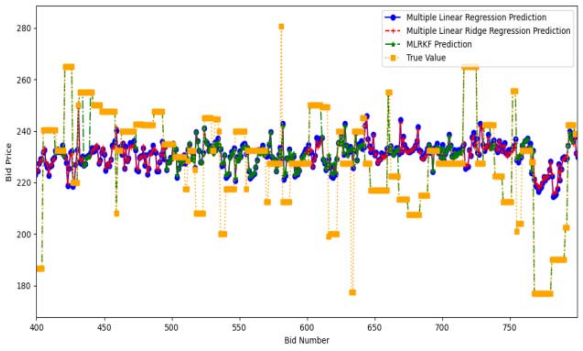
(d) Bid number is between 600 and 800.

Figure 7. Prediction results of regression by MLRP，MLRRP and MLRKFP using dataset1.

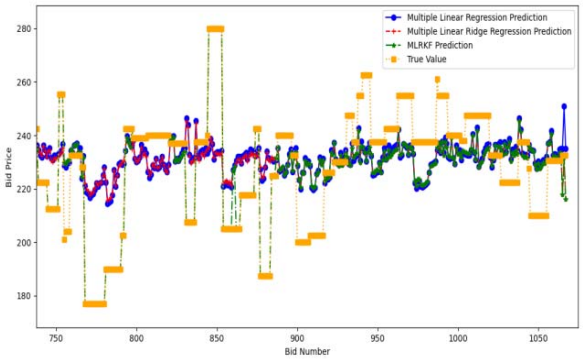
Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.



(a) Bid number is between 0 and 400.



(b) Bid number is between 400 and 750.



(c) Bid number is between 750 and 1050.

Figure 8. Prediction results of regression by MLRP，MLRRP and MLRKFP using dataset2.

The proposed algorithm has the best matching effect with actual auction prices. The results indicate that MLRKF model is a promising model in behavioral bidding tasks.

VII. CONCLUSION

In this paper, a prediction framework named MLRKF is proposed. This prediction framework aims to predict auction end prices and maximize the profit of e-commerce online auction platform. The MLRKF prediction algorithm shows that how to calculate regression weight coefficients and estimate end prices of online auctions. The proposed hybrid algorithm in this paper effectively makes up for shortcomings of the simple multiple linear regression algorithm. MLRP model is a simple model in machine learning, and its prediction accuracy is limited. After simple mathematical transformation and Kalman filter processing, we can complete the prediction more accurately. Meanwhile, the implementation of this algorithm is simple. MLRKF algorithm almost does not increase the calculation of modeling and prediction accuracy has been greatly improved. Therefore, MLRKF algorithm has good performance in model prediction accuracy and modeling calculation.

In the future work, we plan to combine Kalman filter and other machine learning algorithms (e.g. transfer learning and reinforcement learning) to realize online real-time dynamic prediction.

REFERENCES

[1] X. Chen, A. Ghate, A. Tripathi, “Dynamic lot-sizing in sequential online retail auctions,” European Journal of Operational Research, vol.215, no. 1, pp.257-267, 2011.

[2] S.Zhang, W. Jank , G. Shmueli, “Real-time forecasting of online auctions via functional -nearest neighbors,” International Journal of Forecasting,vol.26,no.4, pp.666-683, 2010.

[3] S. Wang, W. Jank, G. Shmueli, “Explaining and forecasting online auction prices and their dynamics using functional data analysis,” Journal of Business & Economic Statistics,vol. 2, pp.144-160, 2008.

[4] P. Kaur, M. Goyal, J. Lu, “Pricing Analysis in Online Auctions Using Clustering and Regression Tree Approach,” in International Workshop on Agents and Data Mining Interaction, Berlin, Heidelberg, 2011.

[5] M. R. Khadge, Manali Rajendra , "Machine learning approach for predicting end price of online auction,” in International Conference on Inventive Computation Technologies , Coimbatore, India,Aug 26-27, 2016.

[6] Yingjie Wang,Yang Gao, Yingshu Li,Xiangrong Tong, “A Worker-selection Incentive Mechanism for Optimizing Platform-centric Mobile Crowdsourcing Systems,” Computer Networks, In press.

[7] D. V. Heijst, R. Potharst, and M. V.Wezel, “A support system for predicting eBay end prices,” Decision Support Syst., vol. 44, pp. 970–982, 2008.

[8] L. Xuefeng, L. Lu, W. Lihua, and Z. Zhao, “Predicting the fifinal prices of online auction items,” Expert Syst. Appl., vol. 31, pp. 542–550, 2006.

[9] D. Lucking Reiley, D. Bryan, N. Prasad, and D. Reeves, “Pennies from eBAy: The determinants of price in online auctions,” J. Ind. Econ., vol. 55, pp. 223–233, 2007.

[10] X. Zhou, W. Liang, S. Huang and M. Fu, “Social Recommendation with Large-Scale Group Decision Making for Cyber-Enabled Online Service,” IEEE Transactions on Computational Social Systems, vol. 6, no. 5, pp. 1073-1082, 2019.

[11] X. Zhou, N.Y. Yen, Q. Jin and T.K. Shih, “Enriching User Search Experience by Mining Social Streams with Heuristic Stones and Associative Ripples,” Multimedia Tools and Applications, vol. 63, no. 1, pp. 129-144, 2013.

[12] X. Zhou and Q. Jin, “A Heuristic Approach to Discovering User Correlations from Organized Social Stream Data,” Multimedia Tools and Applications, vol. 76, no. 9, pp. 11487-11507, 2017.

[13] M.R. Khadge, M.V. Kulkarni, “Machine Learning Approach For Predicting End Price Of Online Auction,” in Proceedings of 2016 International Confeerence on Inventive Computation Technologies(ICICT), Coimbatore, India, Aug.2016,pp. 748–752.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.

[14] R. Gupta, C. Pathak, “A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing,” Procedia Computer Science, vol.36, pp.599-605,2014.

[15] C. Kim, W. Chang, “Logistic regression in sealed-bid auctions with multiple rounds: Application in Korean court auction,” Expert Systems with Application, vol.38, no.4, pp.3098-3115, 2011.

[16] O. Nicolis, M. Diaz, S.K. Sahu, J.C. Marin, “Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile,” Environmentrics, vol.30, no.7, env.2574, 2019.

[17] S. Kumar, "Pricing Algorithms In Online Auctions", International Journal of Advanced Research in Computer Science and Software Engineering, vol.3, no. 6, 2013.

[18] V. Chow, "Predicting auction price of vehicle license plate with deep recurrent neural network", Expert Systems with Applications, vol.142, 2020.

[19] L. Bagadi, G.S. Rao, Ashok, “Firefly,Teaching Learning Based Optimization and Kalman Filter Methods for GPS Receiver Position Estimation,” Procedia Computer Science, vol.143, pp. 892-898, 2018.

[20] F. Zhou, D. Zhong, “Kalman filter method for generating time-series synthetic Landsat images and their uncertainty from Landsat and MODIS observations,” Remote Sensing of Environment, vol.239, pp.111628, 2020.

[21] Kalman, “A new approach to linear filtering and prediction problems,” Journal of Basic Engineering, vol.82, pp.35-45, 1960.

[22] C. Chen, Z. Qin, J.K. Xing, “Prediction Task Assignment of Multi-UAV Approach Based on Consensus,” in Proceedings of International Conference on E-business Technology and Strategy, Ottawa, Canada, 2010.

[23] X. Li, L. Liu, L. Wu, Z. Zhang, “ Predicting the final prices of online auction items,”Expert Systems with Applications,vol.32,no.3,pp.542-550, 2006.

[24] Díaz, Guzmán, Coto, José, Gómez-Aleixandre, Javier, “Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression,” Applied Energy, vol.239, pp.610-625, 2019.

[25] F. Wang, W. Suphamitmongkol, B. Wang, “Advertisement Click-Through Rate Prediction Using Multiple Criteria Linear Programming Regression Mode,” Procedia Computer Science, vol.17, pp.8-3-811, 2013.

[26] X. Li, H. Dong, X. Wang, “Learning to Predict Price based on E-commerce Online Auction Machine,” International Journal of Performability Engineering, vol.14, no.8, pp.1906-1912, 2018.

[27] J. Zhang, L. Edmund, Prater, Ilya Lipkin, “Feedback reviews and bidding in online auctions: An integrated hedonic regression and fuzzy logic expert system approach,” Decision Support Systems, vol.55, no.4, pp.894-902, 2013.

[28] R.D. Cook, “Detection of influential observation in linear regression (reprint from 1977v19 p15-18),” Technometrics, vol 19, no.1, pp. 65-68, 2000.

[29] P.F. Velleman, R.E. Welsch, “Efficient computing of regression diagnostics,” The American Statistician, vol.35, no.4, pp. 234-242, 1981.

[30] X. Song, H. Lu, “Multilinear regression for embedded feature selection with application to fMRI analysis,” in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, Feb. 2017, pp. 2562–2568.

[31] R.M. Hirsch, J.R. Slack, R.A. Smith, “Techniques of trend analysis for monthly water quality data,” Water Resources Research, vol.18, pp. 107-121, 1982.

[32] W. Lutz, W. Sanderson, S. Scherbov, “The end of world population growth,” Nature, vol.412, pp. 543-545, 2001.

[33] M. Shimosaka, T. Tsukiji, H. Wada, K. Tsubouchi, “Predictive population behavior analysis from multiple contexts with multilinear poisson regression,” in Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, Nov.2018, pp. 504–507.

[34] K. Bammann, “Statistical models: Theory and Practice,” Biometrics, vol.62,no.3,pp.940- 9518, 2006.

[35] D.Arjo, “Statistical models: Theory and Practice,” Technometrics, vol.48, no.2, pp.457- 458, 2009.

[36] A.A. Alberto, F. Luis, G.B. Nuria, “Multilinear Weighted Regression (MWE) with Neural Networks for trend prediction,” Applied Soft Computing, vol.82, 2019.

[37] A. Milica, M. Ivan, N. Djordje, “Prediction of Ozone Concentration in Ambient Air Using Multilinear Regression and the Artificial Neural Networks Methods,” Biometrics, vol.62, no.3, pp.3685-3696, 2019.

[38] B.T. Dušan et al., “In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches,” Atmospheric Environment, vol.213, no.15, pp.640-658, 2019.

[39] C.H. Wu, M.Y. Yeh, M.S. Chen, “Predicting winning price in real time bidding with censored data,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1305–1314, 2015.

[40] B.Matzuka, J. Mehlsen, H. Tran , “Using Kalman Filtering to Predict Time-Varying Parameters in a Model Predicting Baroreflex Regulation During Head-Up Tilt,” IEEE Transactions on Biomedical Engineering, vol.62, no.8, pp.1992-2000, 2015 .

[41] C. Liu, G. Zhang, W. Guo, “Kalman Prediction-Based Neighbor Discovery and Its Effect on Routing Protocol in Vehicular Ad Hoc Networks,” IEEE Transaction on Intelligent Transportation, pp.1-11, 2019.

[42] P. Huang, C. Wen, L. Fu, Q. Peng, Z. Li, “ A hybrid model to improve the train running time prediction ability during high-speed railway disruptions,” Safety Science,vol.122,Feb. 2020.

[43] S. D. Liang, M.H. Ma, S.X. He, H. Zhang, “Short-Term Passenger Flow Prediction in Urban Public Transport: Kalman Filtering Combined K-Nearest Neighbor Approach,” IEEE ACCESS, vol.7, pp. 120937-120949, 2019.

[44] W. Chen, W. Zhang, Y.Q. Wu, T.Y. Chen, Z.L Hu, “Short-Term Passenger Flow Prediction in Urban Public Transport: Kalman Filtering Combined K-Nearest Neighbor Approach,” IEEE ACCESS, vol.7, pp. 131653-131662, 2019.

[45] X. Mo, Jing M. Chen, W. Ju, “Optimization of ecosystem model parameters through assimilating eddy covariance flux data with an ensemble Kalman filter,” Ecological Modelling, vol.217(1-2),pp.157-173, 2008.

Authorized licensed use limited to: Columbia University Libraries. Downloaded on July 09,2021 at 02:40:10 UTC from IEEE Xplore. Restrictions apply.