Journal of Construction Engineering and Management

ESTIMATING RESIDUAL VALUE OF HEAVY CONSTRUCTION EQUIPMENT USING ENSEMBLE LEARNING

--Manuscript Draft--

Manuscript Number:	COENG-10335R2
Full Title:	ESTIMATING RESIDUAL VALUE OF HEAVY CONSTRUCTION EQUIPMENT USING ENSEMBLE LEARNING
Manuscript Region of Origin:	SERBIA
Article Type:	Technical Paper
Manuscript Classifications:	Artificial intelligence and machine learning; Business management; Decision making; Quantitative methods
Funding Information:	
Abstract:	Knowing the right moment for the sale of used heavy construction equipment is important information for every construction company. The proposed methodology uses ensemble machine learning techniques to estimate the price (residual value) of used heavy equipment, both present and in the near future. Each machine in the model is represented with four groups of attributes: age and mechanical (describing the machine), and geographical and economic (describing the target market). The research suggests that the ensemble model based on Random Forest, Light Gradient Boosting, and Neural Network members, and Support Vector Regression as a decision unit gives better estimates than the traditional regression or individual machine learning models. The model is built and verified on a large dataset of 500,000 machines, advertised in 50 US states from 1989 till 2012.
Corresponding Author:	Igor Milošević Construction company Dabar Jagodina, SERBIA
Corresponding Author E-Mail:	igornmilosevic@gmail.com
Order of Authors:	Igor Milošević
	Miloš Kovačević, Professor
	Predrag Petronijević, Professor
Suggested Reviewers:	
Opposed Reviewers:	Reviewer #2 from COENG8976R2 Reviewer 2 Reviewer2 This paper is resubmiter. We want to exclude Reviewer #2 from COENG8976R2
Additional Information:	
Question	**Response**
Authors are required to attain permission to re-use content, figures, tables, charts, maps, and photographs for which the authors do not hold copyright. Figures created by the authors but previously published under copyright elsewhere may require permission. For more information see http://ascelibrary.org/doi/abs/10.1061/9780784479018.ch03. All permissions must be uploaded as a permission file in PDF format. Are there any required	No

permissions that have not yet been secured? If yes, please explain in the comment box.	
ASCE does not review manuscripts that are being considered elsewhere to include other ASCE Journals and all conference proceedings (see next question for expanded conference proceeding requirements). Is the article or parts of it being considered for any other publication? If your answer is yes, please explain in the comments box below.	No
Each submission to ASCE must stand on its own and represent significant new information, which may include disproving the work of others. While it is acceptable to build upon one's own work or replicate other's work, it is not appropriate to fragment the research to maximize the number of manuscripts or to submit papers that represent very small incremental changes. ASCE may use tools such as CrossCheck, Duplicate Submission Checks, and Google Scholar to verify that submissions are novel. Does the manuscript constitute incremental work (i.e. restating raw data, models, or conclusions from a previously published study)?	No
Authors are expected to present their papers within the page limitations described in <u><i>Publishing in ASCE Journals: A Guide for Authors</u></i>. Technical papers and Case Studies must not exceed 30 double-spaced manuscript pages, including all figures and tables. Technical notes must not exceed 7 double-spaced manuscript pages. Papers that exceed the limits must be justified. Grossly over-length papers may be returned without review. Does this paper exceed the ASCE length limitations? If yes, please provide justification in the comments box below.	No
All authors listed on the manuscript must have contributed to the study and must approve the current version of the manuscript. Are there any authors on the paper that do not meet these criteria? If the answer is yes, please explain in the comments.	No

Was this paper previously declined or withdrawn from this or another ASCE journal? If so, please provide the previous manuscript number and explain what you have changed in this current version in the comments box below. You may upload a separate response to reviewers if your comments are extensive.	Yes
Please provide the previous manuscript number and explain what you have changed in this current version in the comments box below. You may upload a separate response to reviewers if your comments are extensive. as follow-up to "Was this paper previously declined or withdrawn from this or another ASCE journal? If so, please provide the previous manuscript number and explain what you have changed in this current version in the comments box below. You may upload a separate response to reviewers if your comments are extensive. "	Our article COENG-8976R2 was declined. The editor's final decision was to reject the manuscript with encouragement to resubmit. Last two months, we were working on the new version of the article. The structure of the paper is improved, and we submitted a better version of the research paper.
Companion manuscripts are discouraged as all papers published must be able to stand on their own. Justification must be provided to the editor if an author feels as though the work must be presented in two parts and published simultaneously. There is no guarantee that companions will be reviewed by the same reviewers, which complicates the review process, increases the risk for rejection and potentially lengthens the review time. If this is a companion paper, please indicate the part number and provide the title, authors and manuscript number (if available) for the companion papers along with your detailed justification for the editor in the comments box below. If there is no justification provided, or if there is insufficient justification, the papers will be returned without review.	
If this manuscript is intended as part of a Special Issue or Collection, please provide the Special Collection title and name of the guest editor in the comments box below.	

Papers published in ASCE Journals must make a contribution to the core body of knowledge and to the advancement of the field. Authors must consider how their new knowledge and/or innovations add value to the state of the art and/or state of the practice. Please outline the specific contributions of this research in the comments box. To read more about how the JCEM considers each paper's contributions please see the following Editor's Note.	The proposed methodology uses ensemble machine learning techniques to estimate the price of used heavy equipment, both at present and in the future. The developed Near Future Model enables an owner to calculate the model inputs related to working hours and economic parameters at the time of sale by incorporating the estimate of the short-term engagement of a machine and forecasting the required economic parameters' values. The research suggests that the ensemble model that consists of Random Forest, Light Gradient Boosting and Neural Network members, and Support Vector Regression as a meta learner, gives better prediction results than traditional models. The proposed model could be practically applied in the decision-making process by investors, contractors, or companies engaged in the sale and leasing of heavy construction equipment. The owner of a machine will be able to evaluate its value in a more precise way without hiring experts. The approach makes it possible to improve the overall cost management system of heavy construction equipment.
JCEM strives to continually provide readers with the most relevant data and research. Authors are asked to consider the impact of their research on the global construction community. Papers with a single-region focus should be submitted as a Case Study. Descriptions of a Case Study can be found in <u><i>Publishing in ASCE Journals: A Guide for Authors</u></i>. Does your paper contain research that is focused on a single region?	No
When submitting a new and revised manuscript, authors are asked to include a Data Availability Statement containing one or more of the following statements, with specific items listed as appropriate. Please select any of the statements below that apply to your manuscript. Also, please include the selected statements in a separate "Data Availability Statement" section in your manuscript, directly before the acknowledgements or references. The statement(s) listed in your manuscript should match those you select in your response to this question.	d. Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions.
The flat fee for including color figures in print is $800, regardless of the number of color figures. There is no fee for online only color figures. If you decide to not print figures in color, please ensure that the color figures will also make sense when printed in black-and-white, and remove any reference to color in the text. Only one file is accepted for each figure. Do you intend to pay to include color figures in print? If yes, please indicate which figures in the comments box.	No

Is this article or parts of it already published in print or online in any language? ASCE does not review content already published (see next questions for conference papers and posted theses/dissertations). If your answer is yes, please explain in the comments box below.	No
Has this paper or parts of it been published as a conference proceeding? A conference proceeding may be reviewed for publication only if it has been significantly revised and contains 50% new content. Any content overlap should be reworded and/or properly referenced. If your answer is yes, please explain in the comments box below and be prepared to provide the conference paper.	No
ASCE allows submissions of papers that are based on theses and dissertations so long as the paper has been modified to fit the journal page limits, format, and tailored for the audience. ASCE will consider such papers even if the thesis or dissertation has been posted online provided that the degree-granting institution requires that the thesis or dissertation be posted. Is this paper a derivative of a thesis or dissertation posted or about to be posted on the Internet? If yes, please provide the URL or DOI permalink in the comment box below.	No
If there is anything else you wish to communicate to the editor of the journal, please do so in this box.	We thank the Specialty Editor and the reviewers for the helpful review, which, we believe, resulted in a better presentation of our research. We addressed the remaining issues by rewriting the chapter Related Work and adding several sentences in the chapter Experiments and Discussion.
When submitting a new and revised manuscript, authors are asked to include a Data Availability Statement containing one or more of the following statements, with specific items listed as appropriate. Please select one or more of the statements below that apply to your manuscript. Also, please include the selected statements in a separate "Data Availability Statement" section in your manuscript, directly before the acknowledgments or references. The	a. Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

statement(s) listed in your manuscript should match those you select in your response to this question.	
COPYRIGHT TRANSFER AGREEMENT **I. Authorship Responsibility** To protect the integrity of authorship, only people who have significantly contributed to the research or project and manuscript preparation shall be listed as coauthors. The corresponding author attests to the fact that anyone named as a coauthor has seen the final version of the manuscript and has agreed to its submission for publication. Deceased persons who meet the criteria for coauthorship shall be included, with a footnote reporting date of death. No fictitious name shall be given as an author or coauthor. An author who submits a manuscript for publication accepts responsibility for having properly included all, and only, qualified coauthors. **II. Originality of Content** ASCE respects the copyright ownership of other publishers. ASCE requires authors to obtain permission from the copyright holder to reproduce any material that (1) they did not create themselves and/or (2) has been previously published, to include the authors' own work for which copyright was transferred to an entity other than ASCE. For any figures, tables, or text blocks exceeding 100 words from a journal article or 500 words from a book, written permission from the copyright holder must be obtained and supplied with the submission. Each author has a responsibility to identify materials that require permission by including a citation in the figure or table caption or in extracted text.	None of the exceptions listed above apply.

More information can be found in the guide “Publishing in ASCE Journals: Manuscript Submission and Revision Requirements” (http://ascelibrary.org/doi/pdf/10.1061/9780784479018.ch05). Regardless of acceptance, no manuscript or part of a manuscript will be published by ASCE without proper verification of all necessary permissions to re-use. ASCE accepts no responsibility for verifying permissions provided by the author. Any breach of copyright will result in retraction of the published manuscript.

III. Copyright Transfer

ASCE requires that authors or their agents assign copyright to ASCE for all original content published by ASCE. The author(s) warrant(s) that the above-cited manuscript is the original work of the author(s) and has never been published in its present form.

The undersigned, with the consent of all authors, hereby transfers, to the extent that there is copyright to be transferred, the exclusive copyright interest in the above-cited manuscript (subsequently called the “work”) in this and all subsequent editions of the work (to include closures and errata), and in derivatives, translations, or ancillaries, in English and in foreign translations, in all formats and media of expression now known or later developed, including electronic, to the American Society of Civil Engineers subject to the following:

• The undersigned author and all

coauthors retain the right to revise, adapt, prepare derivative works, present orally, or distribute the work, provided that all such use is for the personal noncommercial benefit of the author(s) and is consistent with any prior contractual agreement between the undersigned and/or coauthors and their employer(s).

• No proprietary right other than copyright is claimed by ASCE.

• This agreement will be rendered null and void if (1) the manuscript is not accepted for publication by ASCE, (2) is withdrawn by the author prior to publication (online or in print), (3) ASCE Open Access is purchased by the author.

• Authors may post a PDF of the ASCE-published version of their work on their employers' ***Intranet*** with password protection. The following statement must appear with the work: "This material may be downloaded for personal use only. Any other use requires prior permission of the American Society of Civil Engineers."

• Authors may deposit the ***final draft*** of their work in an institutional repository or in their funding body's designated archive upon publication in an ASCE Journal, provided the draft contains a link to the published version at ascelibrary.org, and may request public access 12 months after publication. "Final draft" means the version submitted to ASCE after peer review and prior to copyediting or other ASCE production activities; it does not include the copyedited version, the page proof, a PDF, or full-text HTML of the published version.

• Authors may post the ***final draft*** of their work on open, unrestricted Internet sites 12 months after publication in an ASCE Journal, provided the draft contains a link to the published version at ascelibrary.org.

Exceptions to the Copyright Transfer policy exist in the following circumstances. Select the appropriate option below to indicate whether you are claiming an exception:

• **U.S. GOVERNMENT EMPLOYEES**: Work prepared by U.S. Government employees in their official capacities is not subject to copyright in the United States. Such authors must place their work in the public domain, meaning that it can be freely copied, republished, or redistributed. In order for the work to be placed in the public domain, ALL AUTHORS must be official U.S. Government employees. If at least one author is not a U.S. Government employee, copyright must be transferred to ASCE by that author.

• **CROWN GOVERNMENT COPYRIGHT**: Whereby a work is prepared by officers of the Crown Government in their official capacities, the Crown Government reserves its own copyright under national law. If ALL AUTHORS on the manuscript are Crown Government employees, copyright cannot be transferred to ASCE; however, ASCE is given the following nonexclusive rights: (1) to use, print, and/or publish in any language and any format, print and electronic, the above-mentioned work or any part thereof, provided that the name of the author and the Crown Government affiliation is clearly indicated; (2) to grant the same rights to others to print or publish the work; and (3) to collect royalty fees. ALL AUTHORS must be official Crown Government employees in order to claim this exemption in its entirety. If at least one author is not a Crown Government employee, copyright must be transferred to ASCE by that author.

• **WORK-FOR-HIRE**: Privately employed authors who have prepared works in their official capacity as employees must also

transfer copyright to ASCE; however, their employer retains the rights to revise, adapt, prepare derivative works, publish, reprint, reproduce, and distribute the work provided that such use is for the promotion of its business enterprise and does not imply the endorsement of ASCE. In this instance, an authorized agent from the authors' employer must sign the form below. • **U.S. GOVERNMENT CONTRACTORS**: Work prepared by authors under a contract for the U.S. Government (e.g., U.S. Government labs) may or may not be subject to copyright transfer. Authors must refer to their contractor agreement. For works that qualify as U.S. Government works by a contractor, ASCE acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce this work for U.S. Government purposes only. This policy DOES NOT apply to work created with U.S. Government grants.	
Please type your name below to complete the copyright transfer agreement. This will serve as your digital signature. *I, the corresponding author, confirm that the authors listed on the manuscript are aware of their authorship status and qualify to be authors on the manuscript according to the guidelines above.* *I, the corresponding author, confirm that the content, figures, drawings, charts, photographs, and tables in the submitted work are either original work created by the authors listed on the manuscript or work for which permission to re- use has been obtained from the creator.* *I, the corresponding author, acting with consent of all authors listed on the*	Igor Milosevic

manuscript, hereby transfer copyright or claim exemption to transfer copyright of the work as indicated above to the American Society of Civil Engineers.
as follow-up to "**COPYRIGHT TRANSFER AGREEMENT**

I. Authorship Responsibility

To protect the integrity of authorship, only people who have significantly contributed to the research or project and manuscript preparation shall be listed as coauthors. The corresponding author attests to the fact that anyone named as a coauthor has seen the final version of the manuscript and has agreed to its submission for publication. Deceased persons who meet the criteria for coauthorship shall be included, with a footnote reporting date of death. No fictitious name shall be given as an author or coauthor. An author who submits a manuscript for publication accepts responsibility for having properly included all, and only, qualified coauthors.

II. Originality of Content

ASCE respects the copyright ownership of other publishers. ASCE requires authors to obtain permission from the copyright holder to reproduce any material that (1) they did not create themselves and/or (2) has been previously published, to include the authors' own work for which copyright was transferred to an entity other than ASCE. For any figures, tables, or text blocks exceeding 100 words from a journal article or 500 words from a book, written permission from the copyright holder must be obtained and supplied with the submission. Each author has a responsibility to identify materials that require permission by including a citation in the figure or table caption or in extracted text.

More information can be found in the guide “Publishing in ASCE Journals: Manuscript Submission and Revision Requirements” (http://ascelibrary.org/doi/pdf/10.1061/9780784479018.ch05). Regardless of acceptance, no manuscript or part of a manuscript will be published by ASCE without proper verification of all necessary permissions to re-use. ASCE accepts no responsibility for verifying permissions provided by the author. Any breach of copyright will result in retraction of the published manuscript.

III. Copyright Transfer

ASCE requires that authors or their agents assign copyright to ASCE for all original content published by ASCE. The author(s) warrant(s) that the above-cited manuscript is the original work of the author(s) and has never been published in its present form.

The undersigned, with the consent of all authors, hereby transfers, to the extent that there is copyright to be transferred, the exclusive copyright interest in the above-cited manuscript (subsequently called the “work”) in this and all subsequent editions of the work (to include closures and errata), and in derivatives, translations, or ancillaries, in English and in foreign translations, in all formats and media of expression now known or later developed, including electronic, to the American Society of Civil Engineers subject to the following:

• The undersigned author and all coauthors retain the right to revise, adapt, prepare derivative works, present orally, or distribute the work, provided that all such use is for the personal noncommercial benefit of the author(s) and is consistent with any prior contractual agreement between the undersigned and/or coauthors and their employer(s).

• No proprietary right other than copyright is claimed by ASCE.

• This agreement will be rendered null and void if (1) the manuscript is not accepted for publication by ASCE, (2) is withdrawn by the author prior to publication (online or in print), (3) ASCE Open Access is purchased by the author.

• Authors may post a PDF of the ASCE-published version of their work on their employers' ***Intranet*** with password protection. The following statement must appear with the work: "This material may be downloaded for personal use only. Any other use requires prior permission of the American Society of Civil Engineers."

• Authors may deposit the ***final draft*** of their work in an institutional repository or in their funding body's designated archive upon publication in an ASCE Journal, provided the draft contains a link to the published version at ascelibrary.org, and may request public access 12 months after publication. "Final draft" means the version submitted to ASCE after peer review and prior to copyediting or other ASCE production activities; it does not include the copyedited version, the page proof, a PDF, or full-text HTML of the published version.

• Authors may post the ***final draft*** of their work on open, unrestricted Internet sites 12 months after publication in an ASCE Journal, provided the draft contains a link to the published version at ascelibrary.org.

Exceptions to the Copyright Transfer policy exist in the following circumstances. Select the appropriate option below to indicate whether you are claiming an exception:

• **U.S. GOVERNMENT EMPLOYEES**: Work prepared by U.S. Government employees in their official capacities is not subject to copyright in the United States. Such authors must place their work in the public domain, meaning that it can be freely copied, republished, or redistributed. In order for the work to be placed in the public domain, ALL AUTHORS must be official U.S. Government employees. If at least one author is not a U.S. Government employee, copyright must be transferred to ASCE by that author.

• **CROWN GOVERNMENT COPYRIGHT**: Whereby a work is prepared by officers of the Crown Government in their official capacities, the Crown Government reserves its own copyright under national law. If ALL AUTHORS on the manuscript are Crown Government employees, copyright cannot be transferred to ASCE; however, ASCE is given the following nonexclusive rights: (1) to use, print, and/or publish in any language and any format, print and electronic, the above-mentioned work or any part thereof, provided that the name of the author and the Crown Government affiliation is clearly indicated; (2) to grant the same rights to others to print or publish the work; and (3) to collect royalty fees. ALL AUTHORS must be official Crown Government employees in order to claim this exemption in its entirety. If at least one author is not a Crown Government employee, copyright must be transferred to ASCE by that author.

• **WORK-FOR-HIRE**: Privately employed

authors who have prepared works in their official capacity as employees must also transfer copyright to ASCE; however, their employer retains the rights to revise, adapt, prepare derivative works, publish, reprint, reproduce, and distribute the work provided that such use is for the promotion of its business enterprise and does not imply the endorsement of ASCE. In this instance, an authorized agent from the authors' employer must sign the form below. • **U.S. GOVERNMENT CONTRACTORS**: Work prepared by authors under a contract for the U.S. Government (e.g., U.S. Government labs) may or may not be subject to copyright transfer. Authors must refer to their contractor agreement. For works that qualify as U.S. Government works by a contractor, ASCE acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce this work for U.S. Government purposes only. This policy DOES NOT apply to work created with U.S. Government grants."	

ESTIMATING RESIDUAL VALUE OF HEAVY CONSTRUCTION EQUIPMENT USING ENSEMBLE LEARNING

Igor Milošević[1], Miloš Kovačević[2], Predrag Petronijević[3]

[1]Civil Engineer, Construction company Dabar, Vihorska 17, Jagodina, Serbia. Email: igornmilosevic@gmail.com

[2] Professor, Faculty of Civil Engineering, Bulevar kralja Aleksandra 73, Belgrad, Serbia

[3] Professor, Faculty of Civil Engineering, Bulevar kralja Aleksandra 73, Belgrad, Serbia

Abstract

Knowing the right moment for the sale of used heavy construction equipment is important information for every construction company. The proposed methodology uses ensemble machine learning techniques to estimate the price (residual value) of used heavy equipment, both present and in the near future. Each machine in the model is represented with four groups of attributes: age and mechanical (describing the machine), and geographical and economic (describing the target market). The research suggests that the ensemble model based on Random Forest, Light Gradient Boosting, and Neural Network members, and Support Vector Regression as a decision unit gives better estimates than the traditional regression or individual machine learning models. The model is built and verified on a large dataset of 500,000 machines, advertised in 50 US states from 1989 till 2012.

Introduction

The development of construction and the increased complexity of construction projects resulted in the increased engagement of large and expensive construction machinery during project implementation. Buying heavy construction equipment represents a serious investment for every construction company (Pitroda and Chetna, 2015). Over the past decade, in the United States, more than 100 billion dollars have been annually invested on average by companies in the procurement of new and used heavy construction equipment. In 2018, the largest American manufacturer, Caterpillar sold construction machinery worth 23.1 billion dollars (Catepillar, 2019). From the construction management point of view, an owner would like to know the real market value of a machine, to understand how it changes over time, as well as to see what factors and to what extent affect its market value. Higher prices of new equipment have forced a lot of companies to purchase the used machinery. Therefore, both the owner and potential buyer should estimate the market value of used machinery at present and in the near future (from one to two years).

In this paper, the residual value of a used machine is regarded as a function of time and is defined as a price at which the machine can be sold on the market at any given moment. Estimating the residual value of heavy machinery is necessary for calculating the actual cost of performing construction works and for calculating the cost of idle equipment in case of delay claims (Stojadinovic, 2018). The residual value is affected by different types of variables, such as mechanical characteristics, machine condition, market trends, and macroeconomic parameters. The aim of this research is to create an estimation model that considers all these factors and is applicable to different classes of machines. The proposed methodology is based on a machine learning regression model that uses auction web sites as a valuable source of training data. The advertised prices of machines are treated as the best approximations of the unknown selling prices. As opposed to previous studies that are reported in the next section, the model combines individual regression approaches by proposing the usage of the *stacking* ensemble learning technique (Wolpert, 1992).

The proposed model was created (and validated) from a dataset comprising of 500 000 construction machines advertised on numerous US auction sites from 1989 till 2012. As opposed to related work described in the next section, this research utilizes a quantitatively and qualitatively improved data set of diverse construction equipment. It is shown that the proposed ensemble learning technique outperforms traditional estimation approaches (Experiment 2).

Although the approach is tested using data from U.S. auction sites, the methodology for data collection, preprocessing, and model training procedure can be applied to those target markets for which data are available. The advantage of using machine learning techniques is that they can capture a particular market's specifics, which are hidden in the data.

Related research

Unlike project-related data, the construction equipment market offers valuable public information about the characteristics and prices of available mechanization. Estimating the residual value of cars and mechanization was treated in the first printed editions of the Kelley Blue Book, founded in 1926 (James and Waleed, 2005). The age of the equipment and the value of mileage represented the basic variables for determining residual value.

Among the first researchers who studied the residual value, Cubbage attempted to determine the linear dependence between a purchase price and final residual value, claiming that the later varies from 15% to 25% of the initial purchase price (Cubbage, Burgess, and Stokes, 1991). The first significant step forward was made when Cross and Perry conducted a study on the depreciation of agricultural equipment and argued on the shortcomings of previous attempts used to obtain the residual value (Cross and Perry, 1995). Cross and Perry believed that catalog prices could be considered the closest available values that represented the unknown selling prices. However, the authors found that auction prices were one of the best sources of information for estimating real residual values. They observed several predictors such as manufacturer, year of production, size class, condition, operating time, special options, auction type, and region.

In (Unterschultz and Mumey, 1996), the authors considered the impact of changes in technology, quality, and loss in economic value, on the value of heavy equipment. The authors observed the age, hours of use, size, and condition of the equipment. The residual value was calculated by observing the selling price of the equipment that was only one-year-old. In (Kastens, 2002; Lucko and Vorster, 2003), the authors proposed similar empirical formulas for calculating residual value estimates - Vorster and Kastens formula (VK). According to the VK formula (Experiments - Experiment 2), the residual value of a machine is directly proportional to its purchase price and decreases with the square root of its operating hours.

Lucko (2003), Lucko, Anderson-Cook, and Vorster (2006), Lucko, Vorster, and Anderson-Cook (2007), and Lucko and Mitchell (2010), dealt in detail with determining residual values by using linear regression models in which certain input variables were squared (i.e., age) or square rooted (i.e., operating hours). (Lucko, 2003) developed a regression model for estimating the residual value of various types of heavy construction equipment. The predictors were age, manufacturer, condition assessment, geographical area, and certain macroeconomic indicators. (Lucko et al., 2006) advised that the simplest factors should be taken into account while choosing a regression model, i.e., the model should be easy to fit, easy to understand, easy to apply, and easy to justify. Their research hypothesis is that the residual value of the equipment drastically changes under different economic conditions. Lucko et al. (2007) and Mitchell, Hildreth, and Vorster (2011) investigated cumulative values of machine costs to provide a better understanding of the decrease in residual value, depending on the age of the machine.

The application of machine learning (ML) methods in the field of construction project management is gaining in popularity in the last decade. Three papers from the broader context of construction project management that influenced this research are cited: in (Chou and Lin, 2013), the task of early prediction of dispute propensity in public-private partnership projects about public infrastructure services is treated as a classification problem. Authors showed that ensemble techniques provide better prediction accuracy compared to individual classification models. In (Bayzid, Mohamed, and Al-Hussein, 2016), the authors tried to predict the maintenance cost of road construction equipment and showed that regression trees performed better than other nonlinear methods. The ensemble methods are also examined in the most recent study of (Elmousalami, 2020) who analyzed future trends for cost model development in construction engineering and developed a reliable parametric cost model at the conceptual stage of the project.

The first application of ML, to predict residual values on a large dataset of 8589 loaders, was performed in the work of (Fan, et al., 2008). By using the technique of autoregressive decision trees, the authors obtained the estimates of residual values with greater accuracy than by using standard regression models. The single regression tree algorithm provided a good interpretation of the model by using "if-then" analysis. According to them, the entire estimation process can be automated in real-time to follow the auction market changes. In (Zong, 2017), the author observed the manufacturer, machine model, machine age, operating hours, and macroeconomic indicators and compared k-nearest neighborhood, decision tree, and random forest for the task of predicting maintenance cost and a residual value of heavy construction equipment. The study (Milosevic, Petronijevic, and Arizanovic, 2020) established several models based on symbolic regression where input variables were the machine model, age, operating hours, and the inflation index.

The analysis of the previously mentioned work shows the existence of three basic modeling approaches. The first approach uses empirical formulas to estimate residual values based on a machine purchase price and its operating hours (Kastens, 2002; Lucko and Vorster, 2003). This enables easy calculation of residual values, but the estimates are not precise since many important factors that influence residual value, such as mechanical characteristics of a machine (model id, horsepower, hydraulics, drive system, etc.), characteristics of a local market (geographic location, model popularity, the volume of sale, etc.), or macroeconomic parameters (GDP, producer price index, consumer price index, etc.), are not taken into account.

The second (most common) approach involves the creation of a linear regression model using several input variables available to a researcher (Cubbage, Burgess, and Stokes, 1991; Cross and Perry, 1995; Unterschultz and Mumey, 1996; Lucko, 2003; Lucko, Anderson-Cook, and Vorster, 2006; Lucko, Vorster, and Anderson-Cook. 2007; Lucko and Mitchell, 2010). Proposed models use different input variables groups, such as age, mechanical characteristics, characteristics of a local market, and macroeconomic parameters. Nevertheless, none of them uses input variables from all groups. This approach is justified if there exists a linear relationship between the inputs and the residual value. Besides, models built on different machine categories are not transferable because categories show different depreciation types (Cross and Perry, 1995; Fan et al., 2008).

The third (and most recent) approach assumes the nonlinear relationship between input variables and residual values, using nonlinear ML techniques for modeling (Fan et al., 2008; Zong, 2017; Milosevic et al., 2020). The nonlinear models are more accurate than models from two previous approaches, but they require more training data to obtain the desired performance. Nevertheless, small-sized proprietary data sets, containing only one category of machines sold in a short period, were used and analyzed. Existing ML-based models use only a few input variables from all mentioned groups. Although the commonly used mechanical characteristics of machines have a significant impact on residual values, there are many more available on the auction sites that are not exploited when building prediction models. Besides, by monitoring the auction sites over time, one can derive many interesting variables that describe the sales trends in different local markets. In this paper, a model that utilizes as many as possible input variables from the mentioned groups, and a suitable ML technique that can cope with the increased size of inputs, is proposed.

Estimating residual values using the machine learning ensemble approach

Since the residual value of used heavy equipment is treated as a function of time, this research investigates different ML techniques to estimate it both at present and in the near future. The approach assumes the existence of an unknown function g, which maps construction machines to their residual values at a specific point in time. The function g could be approximated with a function f, using a training set of machines described with its characteristics ($\mathbf{x}$) and corresponding residual values (y). Function $y = f(\mathbf{x})$ represents a regression model of residual value. In this research, each machine $\mathbf{x}$ is described as a vector of input attributes grouped into four criteria groups: *Mechanical* (machine class, product size, drive system, etc.), *Age* (number of operating hours, year of production, machine sales date, etc.), *Geographical* (the state where the machine is sold), and *Economic* (PPI, GDP, etc.). When building the model f on a training set, one aims to find $f \approx g$, which will generalize well – it should be capable of predicting residual values from the inputs that describe newly encountered machines.

An ML regression model, which predicts residual values at present and in the near future (next year), is presented in Fig. 1a. The main assumption is that the unknown function g could be inferred from the publicly available auction data originating from specialized web sites. The learning process usually chooses a model f from the preselected family of functions. It then seeks the model-dependent parameters $\mathbf{w}$ ($y = f(\mathbf{x}, \mathbf{w})$) that minimize the difference between the actual and predicted output values on the training data (empirical error). Different ML methods use various error functions, which measure the empirical error, and different approaches for error function minimization with respect to model parameters $\mathbf{w}$.

Ensemble learning

Unlike individual ML methods that learn a mapping f directly from data, the ensemble method constructs a set of mappings and combine their outputs to strengthen the final decision (Zhi-Hua, 2012). In this research, a *stacking* in which several ML methods are trained over the entire data (Wolpert, 1992) is proposed. The structure of the ensemble consists of the basic level models and the decision model (Fig. 1b). Basic level models are trained on the original inputs (machine characteristics vector $\mathbf{x}$). The decision model is trained to map basic level predictions to the final target value y (residual value of $\mathbf{x}$). The stacking aims to minimize the negative impact of input data variation on different learning methods and, at the same time, to increase the overall predictability of the model.

In practice, to configure a good ensemble, two necessary conditions must be met: accuracy and diversity of basic level models (Windeatt and Gholamreza, 2004). Since auction data do not always contain all the information, it is necessary to choose ML methods that can overcome the “missing data” problem (i.e., incomplete vector $\mathbf{x}$). In this research, the selected suitable methods are Random Forest (RF), Light Gradient Boosting (LGB), and Neural Network (NN) – Fig. 1b. The RF (Breiman, 2001) is based on a set of regression trees (Breiman, et al., 1984). It creates a large number of trees, each of which is trained on a random sample of the training set, and searches only on randomly generated subsets of input variables to determine the appropriate split in every node of each tree. RF outputs the averaged prediction of all regression trees. Therefore, it is less sensitive to variations in input data than the predictions of individual trees. Since the trees are less correlated, RF avoids overfitting and reduces the variance of the final model.

LGB (Ke et al., 2017), similar to RF, is a learning technique based on regression trees. It builds a model in iterations by successively adding regression trees to the ensemble, and, like other boosting methods, it improves by reducing the error from a previous iteration. Adding a new tree reduces the error function in the direction of its steepest descent (antigradient). LGB can be successfully combined with RF in an ensemble.

NN is known as a universal functional approximator (Ripley, 1996). In this research, a two-layer feed-forward NN is trained using the backpropagation algorithm, which utilizes gradient descent to minimize the squared error loss function. Here, the error represents the averaged squared difference between the predicted and the real residual values on a training set, and it hopefully decreases with each iteration of training. The process is repeated until the error on a separate validation set starts to increase. The NN method is added to the ensemble since it is commonly used as a regression technique in different scientific fields and contributes to the diversity of the ensemble.

To learn the importance of each basic level model in the ensemble from Fig. 1b, a Support vector regression (SVR) method is used as a decision model (Drucker, Burges, Kaufman, Smola, and Vapnik, 1997). This method is able to construct the regression hyperplane, which is less sensitive to noisy input data than traditional regression methods. SVR uses a kernel function that maps the original input instance into a higher dimensional feature space. It then applies a sort of a linear regression algorithm in the feature space. In this research, a linear kernel was successfully applied, suggesting that there was a linear relationship between the predictions of basic level methods and the final target value.

Since the available data could contain machines advertised on auction sites at different points in time, many ML techniques suffer from the drift problem (Indrė Žliobaitė, 2014). The drift concept represents a change in the relation between the input and the output data over time. However, the appropriate selection of methods can reduce the drift. A motivation for using the ensemble approach, and the particular ML methods, are summarized in Table 1.

Model creation steps

The proposed ensemble model assumes the existence of publicly available auction data, which contains information about advertised prices of construction equipment and their relevant characteristics. Auction websites such as Ritchie Bros, Bidadoo, Equipment Trader, and others represent a valuable source of information about used construction equipment. These websites contain structured information about prices, technical characteristics, age of machines, and additional information such as textual records of machine descriptions that the bidders write in the listings. In the first step of model creation, one must collect the required data from such websites (Fig 2a). There exist a lot of crawling and web scraping services that could help in automating this task. Web scraping is a technique for automated extraction of publicly available information from websites using internet services such as Scrapy, Parsehub, Import.io, and others. The extracted data can be exported in TXT, CSV, HTML, or XLSX formats. Although the scraping requires a sophisticated approach to extract the information from diverse page layouts, it is affordable for companies (i.e., scraping half a million pages usually costs around 200$). The next step assumes the preprocessing of the collected data to remove the obvious errors in descriptions of machines and their prices (i.e., the wrong name of the manufacturer) and fill the missing data (i.e., missing operating hours). There are several strategies to perform data cleansing conducted in the preprocessing step (Fig 2a), which will be discussed in the section Dataset. In the learning process, regression models could benefit from the derived machine characteristics added to the original ones. The derived attributes were added in the features engineering step (Fig. 2a). The original and the derived attributes will be described in detail in section Dataset.

After data acquisition, preprocessing and features engineering, a resulting dataset is used to train the ensemble from Fig. 1b in a two-stage process (Fig 2b). In the first stage, all basic level methods are trained on machines represented as vectors of attributes and corresponding residual values. The decision model (SVR) is trained in the second stage using pairs of values representing the predictions of the basic level models and corresponding residual values. The trained ensemble from Fig. 2b is able to predict residual values at present (the time when the attribute values are observed). To predict residual values in the near future (one or two years ahead), one must transform the time-dependent attributes in the machine representation and then use the trained ensemble (Fig. 2c). Time-dependent attributes, such as machine working hours, or certain macroeconomic parameters that describe the market environment, should be transformed to reflect the machine (and market) state in the near future. The transformation depends on the attribute type and is further described in Experiment 3.

Limitations and assumptions

The main assumption of the research is that the last advertised price is very close to the selling price, and therefore can be considered as the residual value of a machine (Cross and Perry, 1995). Auction sites usually do not contain information about the general condition of machines. This information could only be relevant if there is an independent evaluator who would evaluate individual machines by using the same criteria. Nevertheless, the proposed model takes into account attributes such as machine working hours and the presence or absence of missing parts, which can convey implicit information about the condition of a machine.

The proposed model assumes that mechanical and geographical characteristics do not change over time. On the other hand, future values of macroeconomic indicators should be estimated. Finally, auction sites do not include information about the planned engagement of a machine in the near future. Nevertheless, the owner could incorporate the short-term engagement of the machine in the proposed model (using estimated machine working hours on a project) to obtain better estimates. In this research, we introduce the concept of a modified set of input attributes, which will enable the near-future prediction of residual values (Experiment 3). The limitations and assumptions are summarized in Table 2.

Experiments and discussion

Dataset

The ensemble model was built and validated on separate subsets of 500,000 advertised machines (bulldozers, loaders, trenchers, graders, and excavators), which originated from 50 different US states, from 1989 till 2012. The web data acquisition step from Fig 2a, which included merging data from different sites and deduplication of records using the attributes *Machine ID* and *Model ID*, was done by the company Fast Iron (Fast Iron LLC, 2012). The authors were permitted to use the data acquired by Fast Iron, thus avoiding to crawl the auction sites using the publicly available web scrapping services. Construction machines with a clearly defined machine model, the total number of operating hours, and the year of production between 1989 and 2012 were analyzed. The machines were originally represented with 68 features (Appendix A). The *Sales Price* attribute is assumed as a residual value of a machine, while other descriptive characteristics are treated as independent inputs to the model.

Data preprocessing of the Fast Iron dataset was performed by the authors. The process started with correcting the incorrectly typed or abbreviated values for all attributes (i.e., remove white spaces, Caterpillar instead Cat., John Deer instead John Der, etc.). Missing or illogical attribute values were filled with median (numeric attribute) or mode (categorical attribute) inside the same *Model ID* group if there were enough non-missing values for the attribute (above 75%). Otherwise, the missing values were replaced with the special value of '?'. In order to compare the advertised prices (residual values) from different years and to make a valid regression model, it was necessary to convert all prices to their equivalents in the year 2012 by accounting for different inflation rates:

$$Price_t = Price_{2012} \frac{CPI_t}{CPI_{2012}} \quad (1)$$

where $Price_t$ is the price at the time of the transaction, and CPIs are related to consumer price indexes. According to (Lucko 2011), four macroeconomic parameters correlated with the sales of construction machinery are considered: Consumer Price Index (CPI), Gross Domestic Product (GDP), Producer Price Index (PPI), and Industrial Index Production (INDPROD). Macroeconomic parameters were taken from the U.S. Bureau of Labor Statistics.

Exactly 11 derived attributes were introduced in the features engineering process (Appendix B). The derived attributes should better capture the selling trends for different groups of machines on the auction market. Therefore, the dataset was divided into four-month clusters containing machines of the same model and, for each machine in each cluster, the representation is expanded with several attributes that reflect the trends on the market in the previous time cluster. These are like *Previous Cluster Mean Price*, or different counters for the number of machines sold, how many times the *Model ID* is sold in a state, the number of sales in a given state, and similar. Finally, the *Calendar age* of a machine is calculated as a difference between the *Production Year* and the *Sales Date*.

Statistical analyses of time-dependent attributes and the sales price in the dataset (Table 3) suggest that the *Operating Hours* exhibit a highly skewed distribution – there are a lot of old machines in the dataset whose number of operating hours greatly exceeds the mean value. Standard deviations for *Age* and *Sales Price* indicate that the machines are more spread out in that respect - see Wheel Loader and Track Type Tractors categories. These findings justify the application of the proposed data-driven model since the error made by the eventual averaging approach to the residual value estimation could cost a company lots of money.

All data preprocessing and features engineering tasks were done using MS Excel and Python programming environment by the authors (Milosevic, 2020).

Training the ensemble

To build the ensemble, the available dataset must be separated into two disjunctive sets: 449,186 machines sold before 2012 are treated as a training set; 12,458 machines sold in 2012 are used to evaluate the ensemble model and all other tested models (set S_{2012}). The training set is further divided into sets of machines sold in different periods: $S_{<2011}$ (before 2011), S_{2011_1} (1st quarter of 2011), S_{2011_2} (2nd quarter of 2011), and S_{2011_3} (3rd quarter of 2011).

Stage 1: Training of basic level models

Since the machines were spread over ten years, with certain models appearing and disappearing at different moments in time, a suitable time series training and validation protocol were applied (Hansen and Nelson, 2002). Each basic level method assumes method-dependent hyper-parameters to be selected from the predefined set of values before the final model is trained (Table 4).

Hyper-parameters were selected in a special iterative procedure illustrated in Fig. 3. The optimal hyper-parameters are evaluated after averaging the model performance on three specified validation sets, using a Root Mean Squared Error (Tianfeng and Draxler, 2014).

After finding the optimal hyper-parameters for each basic level model B, the final model for B is trained on the whole training set (all machines sold before 2012).

Stage 2: Training of the decision model

The SVR decision model combines the predictions of basic level models into a final residual value estimate. Since it is easier to learn to weigh predictions of basic models than to learn the mapping between machine characteristics and residual values, the optimal hyper-parameters for Linear SVR were found in only one iteration, using the basic level predictions for the machines sold in 2011 (Fig. 4). After the hyper-parameter C has been found, the decision model was trained on the predictions made on the whole S_{2011} set, and the system was ready for testing.

All training and testing tasks were conducted using the Python library Scikit-Learn (Pedregosa et al.,2011). The total training time for the ensemble was 1058 seconds. A company that develops any data-driven prediction model (including standard residual value regression techniques) will have to spend considerable time to collect and preprocess the needed information. However, spending only two hours training the ensemble model is negligible compared to the data acquisition efforts. If the model is created using internal company resources instead of auction data, it will speed up the process of collecting and preprocessing data. Still, the internal model would be less general, considering the lower number of machine categories, time span, and geographically smaller market.

Experiments

Three different experiments (Fig. 5) were conducted to investigate: what input variables and which base-level ML models are suitable for residual value prediction, and how they compare to each other and to the ensemble model; how the ensemble model compares to traditional methods such as Vorster and Kastens (VK) or linear regression model; the possibility of the ensemble to predict near-future residual values.

All models were tested on the set of machines sold in 2012 (S_{2012}) using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Root Mean Squared Error (RMSE) measures (Tianfeng and Draxler, 2014). These are frequently used performance measures for the evaluation of regression models. Due to the existence of squared terms in the RMSE equation, it is more sensitive to cases in which real and predicted values differ a lot from MAE or MAPE.

Experiment 1: Individual ML models versus the ensemble method

The first experiment aimed to compare the individual ML models with the proposed ensemble method – Table 5. The ensemble achieved the best RMSE of $7997, followed by LGB and RF. Interestingly, NN did not capture the mapping between the input attributes and the residual value very well. An explanation could be that LGB and RF internally use the concept of many learners evolved in iterations (LGB) or in the combination (RF), to form their decisions. They better deal with missing values and have a greater capacity to generalize, while NN could be easily overfitted.

To better understand the applicability of the obtained results, a detailed analysis of the ensemble MAPE error is conducted for certain categories of construction machines (MAE divided with the actual value for each data point and then averaged). The results presented in Table 6 reveal that the ensemble model much better predicts residual values of smaller machines (Backhoe loaders and Skid steer loaders) than the values for bigger construction equipment. This is a direct consequence of the higher standard deviation of advertised prices for these categories (see Table 3).

The importance of decisions of each particular basic level model in the ensemble is shown in Table 7. The ranking follows the results from Table 5 – a more accurate learning method gets more importance in the ensemble, with LGB being the most important. The importance of a basic level model is calculated to be proportional to the increase in the prediction error of the ensemble after the model's predictions were permuted, which should break the influence of a model's outcome on the ensemble outcome. To justify the application of linear SVR, instead of a classical linear method, a simple linear and Ridge regression was tested – Table 8. The best results were obtained in the Linear SVR case. This was expected since the SVR method is more robust to the noisy data and can generalize better.

In the last part of Experiment 1, we performed a recursive feature elimination to determine individual attributes' impact on the ensemble performance (Guyon, Weston, Barnhill, and Vapnik, 2002). RFE fits a model with all attributes and then, in each iteration, removes the weakest attributes and rebuilds the model until the specified number of attributes/iterations is reached. The results obtained by RFE indicated that removing any of the mechanical characteristics decreases the ensemble performance. This suggests that the initial selection of all attributes available from the auction sites was correct because they carry essential information about the machine itself. Fig. 6 shows the ranking of the top 18 most important machine characteristics for the prediction of residual value. The most significant is the derived attribute *Previous Cluster Mean Price*. This finding justifies the derivation of new attributes since they better model sales trends in the auction market. Nevertheless, as previous studies have already shown, the category, the model, and the age of a machine are very important attributes that mainly determine its residual value. Interestingly, our model showed that the macroeconomic parameters (INDPROD) did not affect the estimates significantly. It can be argued that the direct influence of macroeconomic parameters is partially hidden by the sales price trends described with the stronger attribute *Previous Custer Mean Price*.

Experiment 2: Comparing the ensemble with traditional models

Equation (2) was proposed by Vorster and Kastens (Kastens, 2002; Lucko and Vorster, 2003) in an attempt to empirically determine the behavior of a residual value:

$$RV = \frac{K * PP}{\sqrt{\frac{h}{1000}}} \tag{2}$$

RV denotes the residual value of a machine, *K* is an adjustment factor from 0 to 1, with lower values for non-standard machines, *PP* is the purchase price of a machine, and *h* represents the machine working hours. According to Equation (2), the value of the machine rapidly decreases at the beginning of use and then slows down in later years.

In this experiment, the machines from the S_{2012} dataset were divided into four machine categories. The aim was to separate the machines that were different in type and size, as well as being the most numerous in their group of machines. A comparison between the ensemble method and the VK for certain production models is shown in Table 9. Even the ensemble method did not use the purchase price of a machine; it showed significantly greater accuracy of prediction. This expected result derives from the more detailed machine description in the ensemble method (VK uses only purchase price and machine hours). In addition to the VK method, researchers and practitioners commonly use traditional Linear regression models to estimate the residual value of heavy equipment (Lucko, 2011; Lucko et al., 2006). A comparison between the ensemble and the linear regression method is shown in Table 10. The ensemble method exhibited substantially better performance than the linear regression model. This finding suggests the existence of the nonlinear relationship between the machine characteristics and the residual value and justifies the application of the ensemble of nonlinear ML methods. Although more complex to train than the standard linear regression model, our approach exhibits nearly 2300$ better MAE, which becomes even more important when estimating a machine fleet's residual value.

Experiment 3: Testing the ensemble in predicting near-future residual values

The proposed ensemble model should be capable of predicting residual values one year ahead. However, a machine owner must calculate the attributes related to working hours and economic parameters at the time of sale by incorporating the estimate of the short-term engagement of a machine and forecasting the values of the required economic parameters.

According to Fig. 6, the most important economic parameter is the *Industrial Production Index* – INDPROD. The biggest variation in the value of this index was seen in the period of the world economic crisis: from 105.34 (January 2007) to 87.07 (June 2009). The growth of industrial production can be obtained from public sources (Board of Governors of the Federal Reserve System, 2019). Since the other economic parameters are strongly correlated with INDPROD, only this parameter was used to model the economic environment.

To examine the possible economic scenarios in the near future, different test sets were derived from S_{2012} by varying the input values representing a machine's operating hours and INDPROD index. INDPROD took discrete values in the range from –9% (crisis) to +9% (expansion). An assumption is made that, during the crisis, the number of working hours for the next year will decrease by 50% compared to the last year. In the normal scenario, in which INDPROD is between –3% (normal – pessimistic) and +3% (normal – optimistic), there is no change in the number of working hours compared to the last year. Similarly, during the expansion, the working hours for the next year will increase by 50% compared to the last year – the number of projects and the demand for machines will presumably increase. Please note that the previously mentioned percentages are hypothetical and do not follow any economic definitions.

Table 11 shows the performance of the model for different economic scenarios. The results showed that the variation in INPROD increases the prediction error ~ 5% (MAE). Under the normal variation of economic conditions, the model adapts accordingly and does not show a significant change in RMSE and MAE (around 2%). The results are in accordance with the findings from Experiment 1 (Fig. 6), where INDPROD showed a significantly lower impact on the residual value than the main machine characteristics. Therefore, the proposed model is robust enough to be used by the practitioners.

, The accuracy of the model decreases if the residual value is estimated in the distant future (after 2+ years). This statement is confirmed after testing differently trained models on the S_{2012} test set. Suppose that one wants to test a model that predicts two years in advance. In that case, the model is trained on machines sold until 2010. During testing, each machine from 2012 is transformed so that its time-dependent attributes correspond to 2010. The results show that the error increases when predicting several years in advance (from 1 to 7 years) (Fig. 7). The accuracy of the model does not decrease drastically in the distant future, but it decreases rapidly in the first three years. The reason for this unexpected result can be sought in market conditions. After the economic crisis, in 2010 and 2011, there was a decline in sales of machinery and increased dispersion in the range of sales prices, and it was more difficult to estimate the residual value even in the near future. Under normal economic conditions, the growth of RMSE and MAE errors would be more even.

Conclusion

The goal of this research was to build a universally applicable model for the estimation of the residual value of heavy construction equipment. The notion of universality assumes that the model can estimate residual values for different machines classes, to utilize as many as possible relevant types of information that influence residual values, to be transferable on other target markets, and to be able to predict residual values in the near future (next 1-2 years). To enable the applicability of the model on different classes of machines and to cover the majority of input variables that influence residual values, the proposed model is built using the available information from numerous auction web sites. To fulfill the prediction model's transferability on different markets and to successfully predict future trends from a larger amount of input data, a machine learning approach was chosen.

The model assumes that the advertised price of a machine is very close to the unknown selling price (residual value) and that the machine's mechanical characteristics will not change over time. However, the machine's operating hours and the macroeconomic parameters of the market could be estimated and incorporated in the model for the near future prediction.

The main contributions of the research are (1) generating the dataset of nearly half a million machines from the initially obtained Fast Iron data set; (2) proposing the ensemble learning approach for model creation, which is capable of learning the nonlinear mapping between the inputs and the residual value; (3) proposing the model for predicting near-future residual values for different macroeconomic scenarios.

When compiling the dataset, fifteen attributes were derived to improve the modeling of local market trends and macroeconomic environment. The machine learning approach uses convenient regression methods to build a stacking ensemble that better adapts to noisy input attributes and missing data. Experiments suggested that the ensemble model, which appropriately combines Random Forest, Light Gradient Boosting, and Neural Networks, yields better prediction results than the individual ML methods, Vorster and Kastens equation, or widely used linear regression models. The proposed method has shown adaptability to different economic scenarios in the near future, particularly for one year ahead. Hence, the owner of a machine could evaluate the residual value in a more precise way, with the possibility to choose the right moment for selling.

The results of this research could be practically applied in the process of decision making by construction companies or companies engaged in the sale and leasing of heavy construction equipment. The approach makes it possible to improve the overall cost management system of heavy construction equipment. The proposed methodology can be used to build prediction models in related areas of application, such as forecasting the selling price of used vehicles.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bayzid, S. M., Mohamed, Y., and Al-Hussein, M. (2016). Prediction of maintenance cost for road construction equipment: A case study. *Canadian Journal of Civil Engineering*, *43*(5), 480–492. https://doi.org/10.1139/cjce-2014-0500

Board of Governors of the Federal Reserve System (US). (2019). Industrial Production Index (INDPRO). Retrieved from https://fred.stlouisfed.org/series/INDPRO

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, L. A. (1984). *Classification and Regression Trees*. Taylor and Francis,.

Catepillar. (2019). Annual Report 2018. https://doi.org/10.3934/math.2019.1.166

Chou, J. S., and Lin, C. (2013). Predicting disputes in public-private partnership projects: Classification and ensemble models. *Journal of Computing in Civil Engineering*, *27*(1), 51–60. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000197

Cross, T. L., and Perry, G. M. (1995). Depreciation Patterns for Agricultural Machinery. *American Journal of Agricultural Economics*, *77*(1), 194–204. https://doi.org/10.2307/1243901

Cubbage, F. W., Burgess, J. A., and Stokes, B. J. (1991). Cross-sectional estimates of logging equipment resale values. *Forest Products Journal*, (41(10)), 16–22.

Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support Vector Regression Machines. *Advances in Neural Information Processing Systems*, 155–161.

Elmousalami, H. H. (2020, January 1). Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-The-Art Review. *Journal of Construction Engineering and Management*, Vol. 146. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678

Fan, H., AbouRizk, S., Kim, H., and Zaïane, O. (2008). Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model. *Journal of Computing in Civil Engineering*, *22*(3), 181–191. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:3(181)

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Hansen, J., and Nelson, R. (2002). Data mining of time series using stacked generalizers. *Neurocomputing*, *43*, 173–184.

Indrė Žliobaitė. (2014). Learning under Concept Drift: an Overview. In *Faculty of Mathematics and Informatics, Vilnius University, Lithuania*.

James, W., and Waleed, M. (2005). Adverse Selection and Reputation Systems inOnline

Auctions: Evidence fom eBay Motors. *International Conference on Information Systems*, 847–858.

Kastens, T. H. (2002). *An Owning and Operating Cost Calculator Model for Construction Equipment. Final Exam submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirement for the degree of Master of Science in Civil*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient BoostingDecision Tree. *Proceedings of Neural Information Processing Systems*.

Lucko, G. (2003). A Statistical Analysis and Model of the Residual Value of Different Types of Heavy Construction Equipment. *Http://Scholar.Lib.vt.Edu/Theses/Available/Etd-12032003-122642*.

Lucko, G. (2011). Modeling the Residual Market Value of Construction Equipment under Changed Economic Conditions. *Journal of Construction Engineering and Management*, *137*(10), 806–816. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000279

Lucko, G., Anderson-Cook, C. M., and Vorster, M. C. (2006). Statistical Considerations for Predicting Residual Value of Heavy Equipment. *Journal of Construction Engineering and Management*, *132*(7), 723–732. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(723)

Lucko, G., and Mitchell, Z. W. (2010). Quantitative Research: Preparation of Incongruous Economic Data Sets for Archival Data Analysis. *Journal of Construction Engineering and Management*, *136*(1), 49–57. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000078

Lucko, G., and Vorster, M. C. (2003). Predicting the Residual Value of Heavy Construction

Equipment. *Towards a Vision for Information Technology in Civil Engineering*, 1–11. https://doi.org/10.1061/40704(2003)49

Lucko, G., Vorster, M. C., and Anderson-Cook, C. M. (2007). Unknown Element of Owning Costs—Impact of Residual Value. *Journal of Construction Engineering and Management*, *133*(1), 3–9. https://doi.org/10.1061/(ASCE)0733-9364(2007)133:1(3)

Milosevic, I. (2020). OSF Project. https://doi.org/DOI 10.17605/OSF.IO/VMBR2

Milosevic, I., Petronijevic, P., and Arizanovic, D. (2020). Determination of residual value of construction machinery based on machine age. *Gradevinar*. https://doi.org/https://doi.org/10.14256/JCE.1285.2015

Mitchell, Z., Hildreth, J., and Vorster, M. (2011). Using the Cumulative Cost Model to Forecast Equipment Repair Costs: Two Different Methodologies. *Journal of Construction Engineering and Management*, *137*(10), 817–822. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000302

Pitroda, J., and Chetna, V. (2015). A critical literature review on comparitive analysis of construction equipments –rent and buy. *Journal of International Academic Research for Multidisciplinary*, (2320–5083), 130–141.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*.

Scholz, M., and Klinkenberg, R. (2005). An Ensemble Classifer for Drifting Concepts. *Proceedings of the Second International Workshop on Knowledge Discovery in Data Streams*.

Tang, F., and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining*, *10*(6), 363–377.

Tianfeng, C., and Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error

(MAE)?– Arguments against avoiding RMSE in the literatur. *Geoscientific Model Development*, *7*(3), 1247–1250.

Unterschultz, J., and Mumey, G. (1996). Reducing Investment Risk in Tractors and Combines with Improved Terminal Asset Value Forecasts. *Canadian Journal of Agricultural Economics*, *44(3)*, 295–309.

Windeatt, T., and Gholamreza, A. (2004). Decision tree simplification for classifier ensembles. *International Journal of Pattern Recognition and Artificial Intelligence*, 749–776.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*(2).

Zhi-Hua, Z. (2012). *Ensemble Methods - Foundations and Algorithms*.

Zong, Y. (2017). *Maintenance Cost and Residual Value Prediction of Heavy Construction Equipment*.

Table 1: How to overcome the problems that appear in the ML modeling process.

Problem	**What to do**	**Literature**
Missing machine data (missing attributes in **x**)	Use of Random Forest, the introduction of new attributes (feature engineering)	(Tang and Ishwaran, 2017)
Drift concept	Ensemble models, regular upgrade of database and repeated machine learning	(Indrė Žliobaitė, 2014; Scholz and Klinkenberg, 2005)

Economic parameters unpredictably change over time	Gradient Boosting methods, ensemble models	(Scholz and Klinkenberg, 2005)

Table 2: Limitations and assumptions

Limitation	**Assumption**
Advertised prices differ from real selling prices	The last advertised price is very close to the selling price
There is no information about the general condition of machines	Machine working hours or the presence or absence of missing parts carry implicit information about the condition of a machine
Owners often do not enter all machine characteristics	Noise and redundancy of the data must be appropriately handled (input data transformation, ensemble methods, Table 1).
Data sets do not include information about the planned engagement of a machine	One must incorporate the estimated short-term engagement of the machine when predicting near future residual values.

Table 3: Mean and standard deviation of the *Operation Hours*, *Age* and *Sales Price*

Machine Type	Operating Hours (h)		Age (year)		Sales price ($)	
	Mean	Std.	Mean	Std.	Mean	Std.

Skid Steer Loaders	2289	29255	9	4	11425	3616
Track Excavators	4172	29781	11	5	40594	25214
Track Type Tractors	3295	26107	15	9	39794	25287
Wheel Loader	4347	27344	15	9	42271	23439
All machines	**3409**	**26625**	**13**	**8**	**34824**	**24961**

Table 4: Hyper-parameters for different models (SVR is used only in the decision level). The names of parameters are taken from the Scikit-Learn library (Pedregosa et al.,2011)

LGB	RF	NN	Linear SVR
max_depth	n_estimators	num_neurons	C
num_leaves	max_features	num_hidden_layers	
learning_rate	min_samples_leaf		
feature_fraction			
bagging_fraction			

Table 5: Comparison of individual ML methods and the proposed ensemble.

	LGB	NN	RF	Ensemble
RMSE ($)	8139	11005	8528	7977
MAE ($)	5452	7251	5667	5359

Table 6: Comparing the performance of the ensemble model between machine categories: Wheel loader (WL), Skid steer loader (SSL), Track excavator (TE), Backhoe loader (BL), Motograder (MG), Track Type Tractor (TTT).

	WL	SSL	TE	BL	MG	TTT
MAE($)	7494	1659	6649	2861	7260	7092
MAPE (%)	20.8	4.8	18.1	8.3	19.7	19.2

Table 7: The relative importance of basic level methods in the ensemble (bigger values – higher importance).

Individual method	Impact of each method
RF predictions	0.364
NN predictions	0.067
LGB predictions	0.569

Table 8: Linear SVR, Linear regression and Ridge regression comparison.

	Linear SVR	Linear Regression	Ridge
RMSE ($)	7977	8442	8423
MAE ($)	5359	5689	5679

Table 9: Comparing RMSE and MAE for the ensemble and VK models. A – Backhoe Loader - 14.0 to 15.0 Ft Standard Digging Depth 310G B – Motor grader - 145.0 to 170.0 Horsepower 140G, C – Skid Steer Loader 763 - 1351.0 to 1601.0 Lb Operating Capacity, D – Track Type Tractor, Dozer D8K - 260.0 Horsepower.

	Backhoe Loader A		*Motor grader B*		*Skid Steer Loader C*		*Tractor Dozer D*	
	Ensemble	VK	Ensemble	VK	Ensemble	VK	Ensemble	VK
RMSE ($)	4065	6426	11392	174599	2078	8046	14168	238368

MAE ($)	3161	5158	7361	52325	1546	7551	10797	93178

Table 10: Comparing the ensemble model to traditional linear regression.

	Ensemble	Linear egression
RMSE ($)	7977	11825
MAE ($)	5359	7613

Table 11: Predicting near future residual values in different economic scenarios.

Economic Scenario	INDPROD	Increase in Machine Working Hours per year (%)	Ensemble RMSE ($)	Ensemble MAE ($)
Crisis	-9%	– 50%	8320	5669
Normal – pessimistic	-3%	0%	8213	5517
Normal	0	0%	8042	5426
Normal – optimistic	+3%	0%	8317	5598
Expansion	+9%	+50%	8324	5621

Fig.1

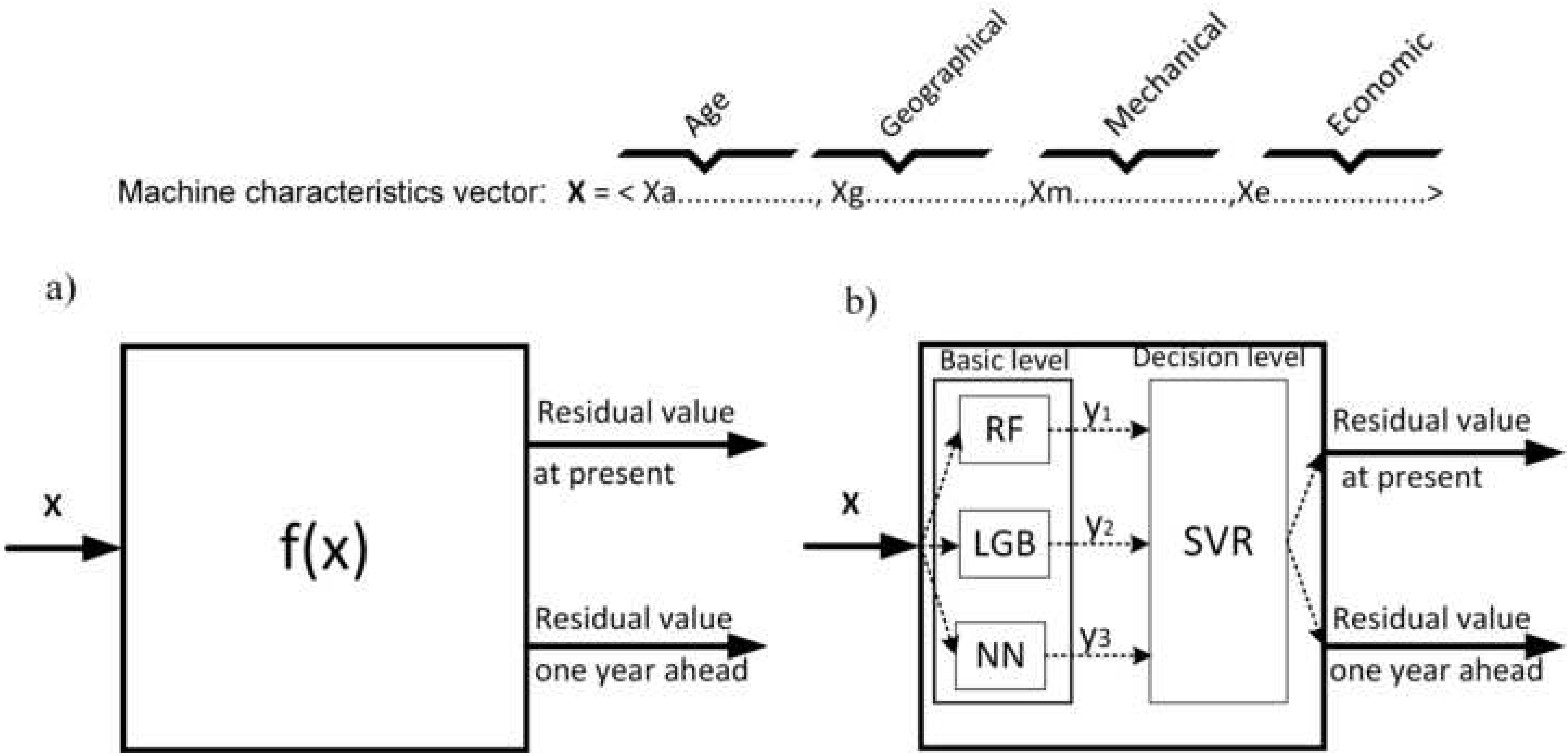
Age
Geographical
Mechanical
Economic
Machine characteristics vector: X = < Xa................, Xg.................,Xm.................,Xe.................>
a)
x
f(x)
Residual value
at present
Residual value
one year ahead
b)
Basic level
Decision level
x
RF
LGB
NN
y1
y2
y3
SVR
Residual value
at present
Residual value
one year ahead

Fig.2

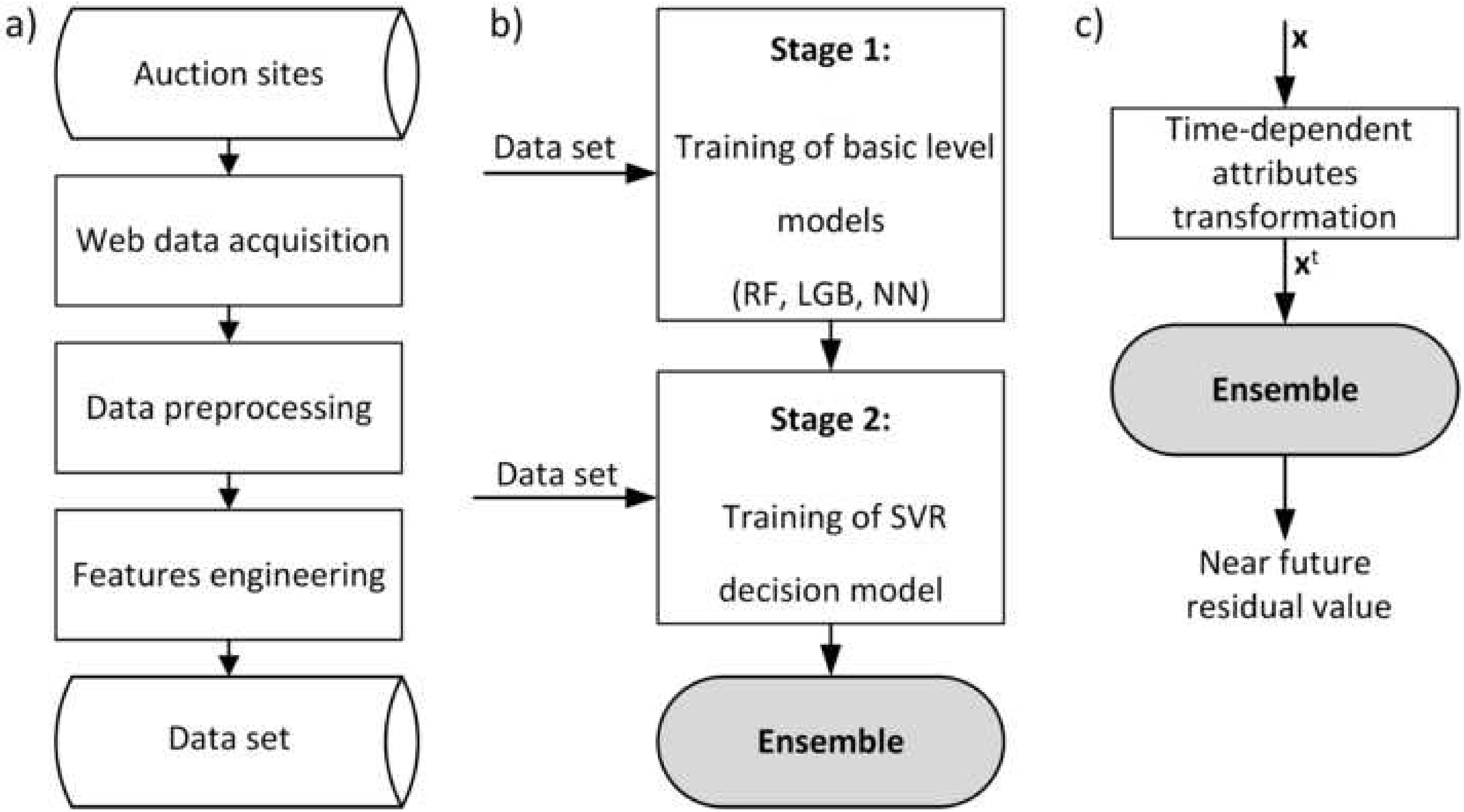
a)
Auction sites
Web data acquisition
Data preprocessing
Features engineering
Data set
b)
Data set
Stage 1:
Training of basic level models
(RF, LGB, NN)
Data set
Stage 2:
Training of SVR decision model
Ensemble
c)
x
Time-dependent attributes transformation
xt
Ensemble
Near future residual value

Fig.3

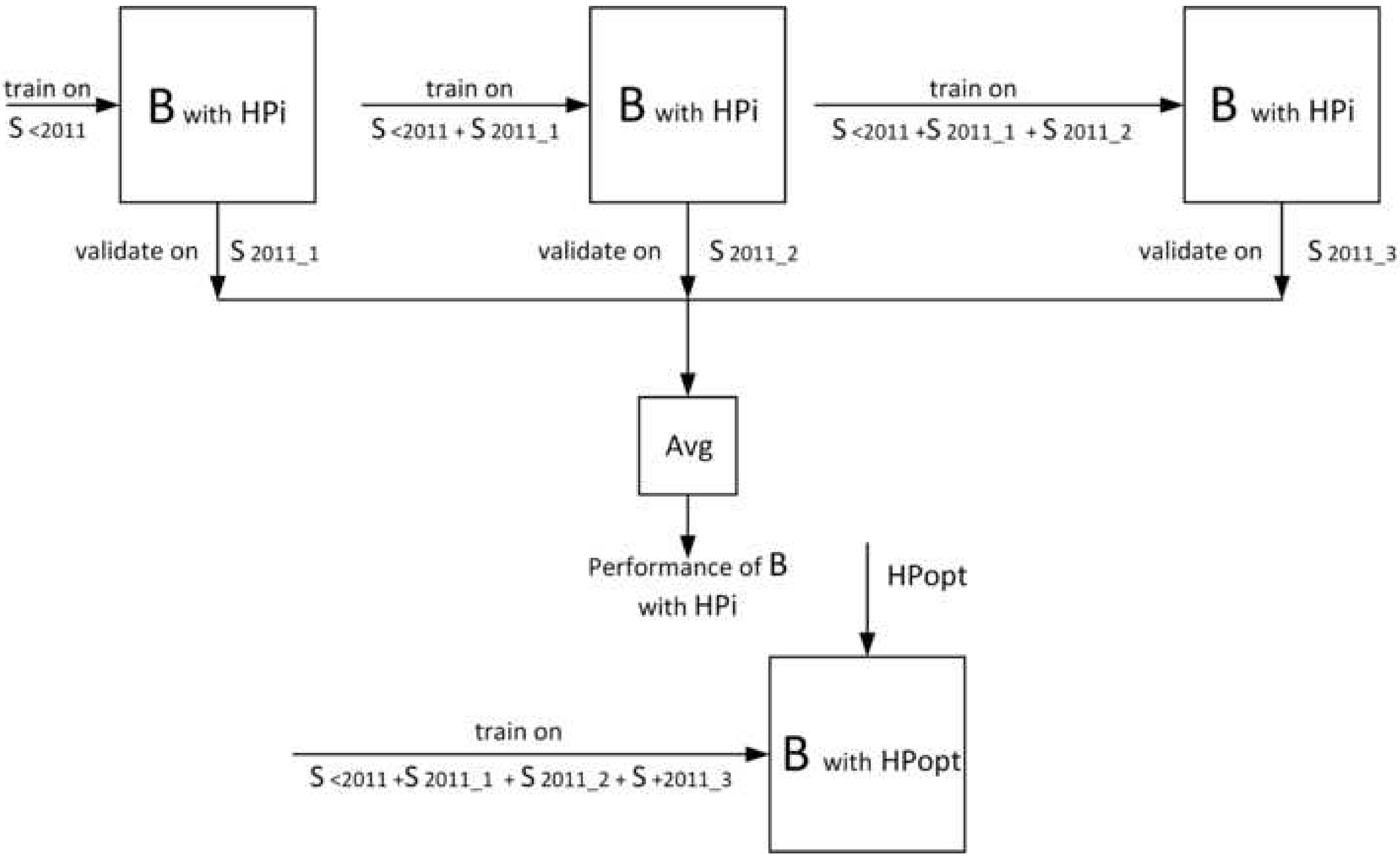
train on
S <2011
B with HPi
train on
S <2011 + S 2011_1
B with HPi
train on
S <2011 +S 2011_1 + S 2011_2
B with HPi
validate on
S 2011_1
validate on
S 2011_2
validate on
S 2011_3
Avg
Performance of B with HPi
HPopt
train on
S <2011 +S 2011_1 + S 2011_2 + S +2011_3
B with HPopt

Fig.4

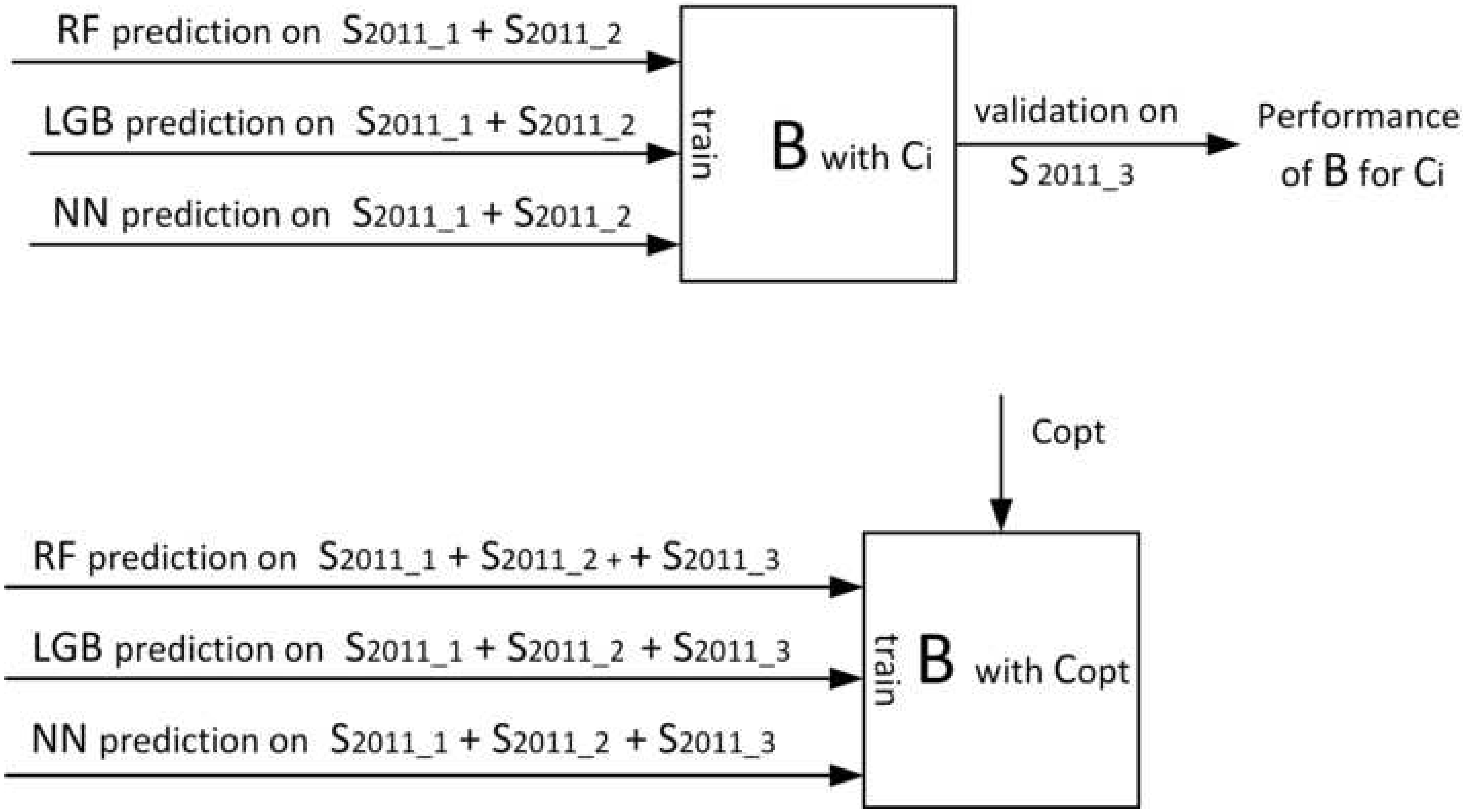
RF prediction on S2011_1 + S2011_2
LGB prediction on S2011_1 + S2011_2
NN prediction on S2011_1 + S2011_2
train
B with Ci
validation on
S 2011_3
Performance
of B for Ci
Copt
RF prediction on S2011_1 + S2011_2 + + S2011_3
LGB prediction on S2011_1 + S2011_2 + S2011_3
NN prediction on S2011_1 + S2011_2 + S2011_3
train
B with Copt

Fig.5

Click here to access/download;Figure;Fig.5.tiff

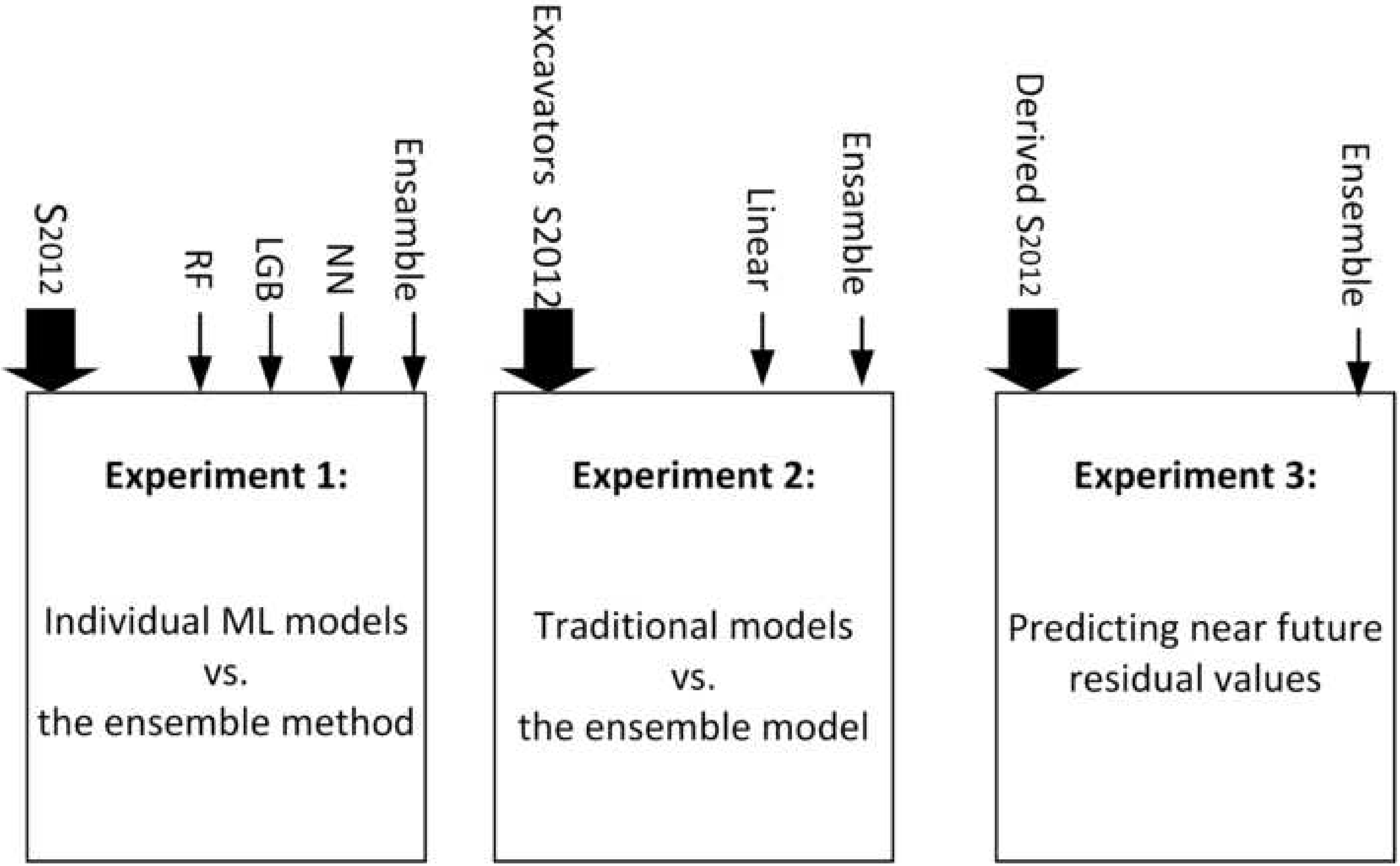

Fig.7

Click here to access/download;Figure;Fig.7.tiff

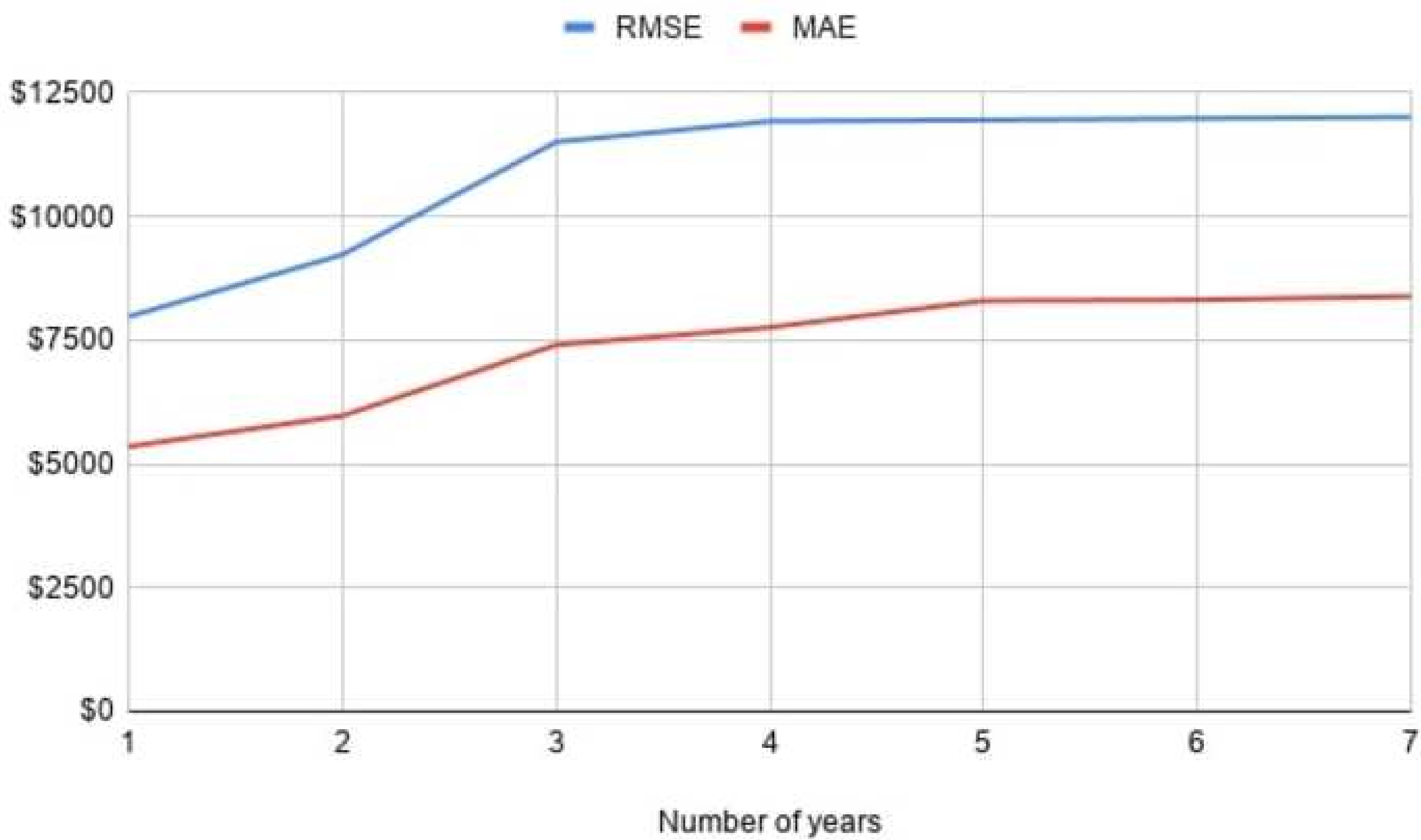

Fig.6

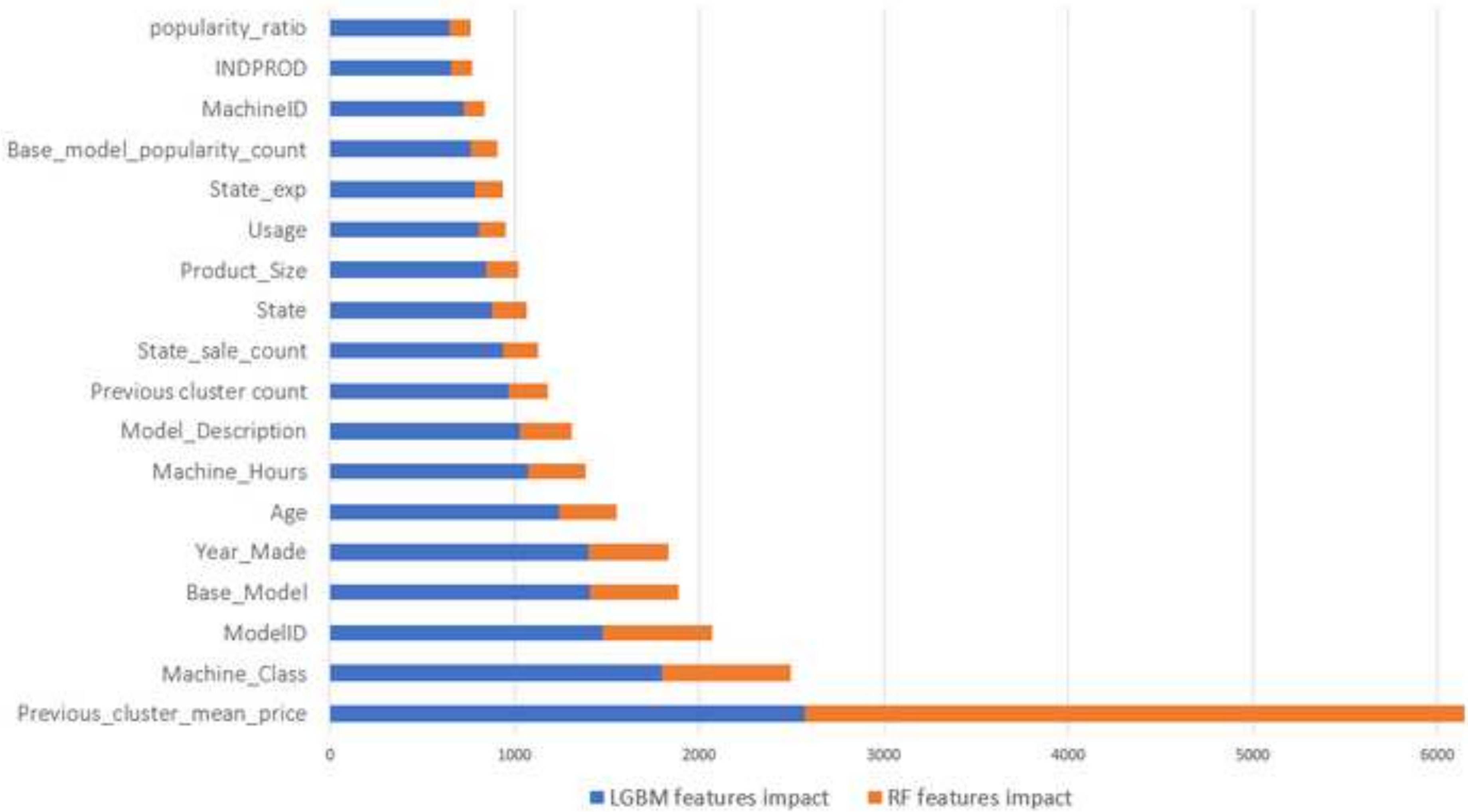
popularity_ratio
INDPROD
MachineID
Base_model_popularity_count
State_exp
Usage
Product_Size
State
State_sale_count
Previous cluster count
Model_Description
Machine_Hours
Age
Year_Made
Base_Model
ModelID
Machine_Class
Previous_cluster_mean_price
0
1000
2000
3000
4000
5000
6000
LGBM features impact
RF features impact

Fig.1. Model representation - Basic view (a): Machine characteristics **x** are mapped into desired residual values $y = f(\mathbf{x})$. The model predicts the present and the near-future residual value of **x**. Ensemble view (b): Inputs are fed into *n* basic level models RG, LGB, NN. Their predictions are combined using a decision model SVR to form the final residual value prediction.

Fig.2. Important steps in the creation of model. (a) Data set creation steps. (b) Ensemble training in a two stage process. (c) Ensemble predicts the near future residual value of machine **x** after transforming its time-dependent attributes (i.e. machine age is incremented). Vector $\mathbf{x}^{\mathbf{t}}$ denotes the transformed representation of the machine, related to the near future.

Fig.3. Each basic level model B (RF, LGB, NN) is trained on all machines sold before 2012. The optimal set of hyper-parameters (HPopt) was chosen after three iterations of training and validation on the specified sets. Model performance, under a fixed set of hyper-parameters, is averaged. An optimal set yields best averaged model performance (minimal RMSE on a validation set).

Fig.4. SVR decision model is trained on all machines sold in 2011. The optimal hyper-parameter for C (Copt) was chosen after training and validation on the specified sets of basic level predictions. Here, each machine from sets S_{2011_1} and S_{2011_2} is represented as a triple of its predicted residual values.

Fig.5. Experiments performed in the research: wide arrows indicate test sets used to evaluate model performance; line arrows denote models.

Fig.6. Most important attributes: the importance of an attribute is proportional to the increase in the prediction error (axis values represent sales price errors) of the model after the attribute's

values were permuted, which should break the relationship between the attribute and the true outcome.

Fig.7. Growth of MAE and RMSE errors due to residual value estimation in the distant future. The number of years is marked in the range from 1 (near future), to 7 (distant future).

APPENDIX A

Click here to access/download
Supplemental Materials File
Appendix A.docx

We thank the Specialty Editor and the reviewers for the helpful review, which, we believe, resulted in a better presentation of our research. We addressed the remaining issues by rewriting the chapter *Related Work* and adding several sentences in the chapter *Experiments and Discussion*. The answers to particular remarks are given below:

Reviewer 2:

Remarks	Answer
The narratives in lines 271-279 and the appendices A and B are not sufficient to describe the feature engineering process. Feature engineering is not just about what features you used but also why do you use them. The authors should demonstrate the rationale for the initial selection of the features. This step is critical for establishing any prediction model.	As we have explained in an updated *Related research* section (lines 120-151), analysis of previous research has shown that none of the earlier studies use all attributes available on auction sites in determining the residual value. We decided to test all of them and see how and to what extent they affect the residual value. In the last part of *Experiment 1,* we performed a recursive feature elimination to determine individual attributes' impact on the ensemble performance and showed that removing any mechanical characteristics decreases the ensemble performance 391 - 398. This analysis resulted in the ranking showed in Figure 6.
I suggest the authors putting the two appendices to the OSF project created for this paper. The information in them is important and should be available for the readers. The Journal of Construction Engineering and Management may not publish the appendices.	Appendices are now presented in the OSF project.
The authors should pay attention to the style of written English. There are some vocabulary and usage of spoken English in the current manuscript. For example, the authors should use "such as" instead of "like." The manuscript should be proofread to avoid this kind of informal usage.	Proofreading and grammar check was done, and corrections were made.

Reviewer 3:

Remarks	Answer
Thank you for addressing the comments and improving the paper. There are a few points needed to be addressed before publishing the paper as follows. It is essential to develop Related Research logically. As a suggestion, the authors could first provide all contents of historical development associated with the residual value estimation of construction equipment, thereafter explaining the essence, advantages, and disadvantages of prior approaches applied for solving existing problems.	The *Related research* chapter is historically rearranged based on the Reviewer suggestion. In the first, historical review part of the Related research chapter, for each study, the used predictors were listed (lines 57-119) Based on the analysis of the related research, in the second part of this chapter, three methodological approaches are listed, and then the advantages and disadvantages of these approaches are presented (120-151).
The paper needs a complete proofreading. There are several grammatical and structural problems while developing the paper. For instance, in line#74, "was" should change to "were". Also, it seems that, in line#88, "to" should change to "on".	Proofreading and grammar check was done, and corrections were made.