USING DATA-SCIENCE MODELS TO PREDICT TECHNOLOGICAL FACTORS AFFECTING THE MECHANICAL PROPERTIES OF FLAT PRODUCTS

Panagiotis Sismanis

Sidenor Steel Industry S.A.,
33, Amaroussiou-Halandriou St.
GR-15125 Maroussi, Athens,
Greece
E-mail: psismanis@sidenor.vionet.gr

Received 13 August 2016
Accepted 16 December 2016

ABSTRACT

In recent years, the stringent specifications lead to the need of more strict regulations within the plants in order to provide products that conform to the market needs. This requires a better appraisal of the mechanisms that influence the demanded material properties in order to set up the appropriate rules. For the production of steel plates for example, a complete steelmaking route from scrap melting, refining, casting, till rolling is required. A great deal of information linked to technological parameters involved in the process is already stored in the automation systems within a plant. With the deployment of machine learning algorithms specific mechanical properties of the final products can be related with these factors and two major results can be obtained:

The most important technological parameters that influence the properties under investigation are deduced;

A supervised model that predicts a mechanical property upon a set of input data can be derived within a measurable statistical error.

In this work, two mechanical properties for produced plates, the tensile strength (Rm), and the yield stress (Re), were analyzed with respect to 33 independent parameters representing salient features in the whole production process. Three data science models, the deep learning, the distributed random forest, and the gradient boosting method were deployed for securing the validation of the results. Attention is drawn upon those technological parameters that are top selected in all three models. Actual and predicted properties values are also presented.

Keywords: data science models, technological parameters, flat products, predicted properties.

INTRODUCTION

The need for a better appraisal of the technological parameters involved that influence the demanded material properties in order to set up the appropriate rules is of paramount importance. However, depending upon the number and complexity of processes involved, it is sometimes a very difficult task to figure out all the technological factors and parameters that influence specific material properties. In the steelmaking sector for example, and specifically in the domain of flat products it takes a great number of high temperature processes in order to come up with a specific product grade. One has to start from the meltshop where scrap is melted in electric arc furnaces (EAFs) and tapped into ladles in which liquid steel is processed in the secondary metallurgy stations (LFs, and in special cases in VDs - vacuum tank degassers), where the specific chemical analysis of the target grade is achieved, and finally casted in big slabs; these slabs are left to cool down and then are reheated in reheating furnaces in order to be rolled under very strict rolling schedules to the final plate dimensions. According to product specifications test-pieces from the final plates are taken and examined for conformity according to specific mechanical properties and material soundness. It is easily realized the negative impact upon cost in case that a product batch fails in one or more properties. In addition to this, potential customer dissatisfaction is very likely to occur in case product supply cannot match the signed deadlines.

It is about a decade that versatile computational tools that apply statistics and can be used for decision making are given away through the internet. Machine learning algorithms have emerged that are capable of handling and manipulating huge numbers of data giving rise to the appearance of data science. The technological parameters treated either as predictors or response variables can be implicitly linked into models with a very well defined statistical error. In this way, supervised models can be developed that can supply a further insight into the phenomena that take place in real practice. In the present world of electronics, automation systems are not only ubiquitous in the plants but collect a great deal of information in real time. Level-I type of data are collected almost every second in industrial applications. Level-II type of information is the result of Level-I data aggregates. In case that Level-II is not installed yet, data aggregates can be produced from Level-I data collection over a specific period. Consequently, a suitable number of predictors and response variables are selected for a relatively large number of processed data. The predictors are in general all the technological parameters that influence the response variables, the latter being the mechanical properties under investigation. Once the data are collected, data frames are constructed with columns being the predictors and the response variable under investigation. After some proper data manipulation the data frames are imported into the machine learning algorithms and an attempt is made in order to come up with supervised models capable of predicting the values of the response variables under various input predictor values. In addition to this, a list of the most important parameters that influence the property under investigation (i.e. response variable) is derived, as well. This is of paramount importance as based on these technological parameters a new set of internal rules can be deduced that will support the control of the property under question to the maximum level.

In this work, a set of data were collected in a two-month period from three different places inside the Stomana plant. Specifically, data were collected from the meltshop, the plate mill, and the quality control with respect to two fundamental material properties, the tensile strength (Rm), and the yield stress (Re) of produced plates. PLC generated data were collected from the automation system of the plate mill; salient features like casting speed, casting temperature, chemical analysis, active oxygen content at EAF tapping, etc, were collected from the meltshop automation systems, as well. The main task was to deduce models that can identify the most important features that affect Rm and Re, and to predict these two properties to a measurable extend within statistical error. In Stomana, a relatively large number of grades are produced for flat products, so it was decided to restrict the analysis only for the S355-based grade products, which comprise the biggest percentage.

Three machine learning algorithms (models) were deployed. The deep learning (DL) [1], the distributed random forest (DRF) [2], and the gradient boosting method (GBM) [3] models were put into practice. The H2O Flow package [4] was used either directly implemented in a locally created cluster or via the R language environment. This package can be downloaded for free from the internet and it is for the moment extensively used in about ten-thousand companies worldwide becoming in such a way a standard. Deep learning is a branch of machine learning where a multilayered (deep) architecture is used to map the relations between inputs or observed features and the outcome (response variable). This deep architecture makes deep learning particularly suitable for handling a large number of variables and allows deep learning to generate features as part of the overall learning algorithm. It is based on neural networks that contain a series of neurons, or nodes, which are interconnected and process input. The connections between neurons are weighted, with these weights based on the function being used and learned from the data. Activation in one set of neurons and the weights (adaptively learned from the data) may then feed into other neurons, and the activation of some final neurons is the prediction [5]. A GBM is an ensemble of either regression or classification tree models. Both are forward-learning ensemble methods that obtain predictive results using gradually improved estimations. Boosting is a flexible nonlinear regression procedure that helps improve the accuracy of trees. Weak classification algorithms are sequentially applied to the incrementally changed data to create a series of decision trees, producing an ensemble of weak prediction models.

While boosting trees increases their accuracy, it also decreases speed and user interpretability. The gradient boosting method generalizes tree boosting to minimize these drawbacks [3]. Finally, the distributed random forest (DRF) is a variation of a general technique called ensemble learning. An ensemble model is composed of the combination of several smaller simple models (often small decision trees). The random forest approach tries to de-correlate the trees by randomizing the set of variables that each tree is allowed to use. The final ensemble of trees is then bagged to make the random forest predictions [6]. Although the whole description sounds a bit too technical, it is the proper call of the corresponding functions (models) that does the job.

EXPERIMENTAL

In total, 1147 cases (rows) of data were collected with 33 independent technological parameters (predictors) influencing the two depended properties (Rm & Re); in this way, more than forty-thousand values were collected and processed. It is understandable that this large number of values is still prone to real world error. It is anticipated that this error is relatively small, of the order of less than 3 %. In industrial conditions, this type of error still larks almost anywhere. The software was run in a DELL Alienware laptop, with the Intel i7-6700HQ CPU (8 cores) @ 2.60GHz, 16GB RAM, running under the 64-bit operating system Windows 10 Professional. At first, a cluster was generated by Java-Virtual-Machine 64-bit-software in which the used memory size, the number of CPU-cores, and the H_2O connection was created and established. Then the set of data (data frame) was imported into the cluster. Each data frame (with 1147 rows by 34 columns for Rm –and one more similar data frame for Re-) was split in two data frames, randomly: the training data frame consisted of the 75 % of data, and the validation data frame consisted of the rest 25 % of the data. The models were trained with the 75 % of the data. Consequently, the derived implicit models are based on the training data; these models are then tested (validated) on the rest 25 % data, generating supervised models with a measurable statistical error. Two types of running programs were executed per model. In the first part of the analysis a grid search was performed in order to deduce the proper tuning parameters in order for the specific model to minimize to a more-or-less extent the validation mean-squared error (MSE). In the second part, the model was executed with the most appropriately selected tuning parameters in order to generate the final supervised model. It should be noted that the overall derivation of a supervised model is a time-consuming process. Even for the great computational capacity of the laptop used in this study a supervised model normally took more than five hours of computing time to conclude.

RESULTS AND DISCUSSION

Tensile strength (Rm) results

As stated above, the training data are used to train the deployed algorithm and come up with a supervised model but the validation data are used for its final acceptance. The mean-squared error (MSE) is a computed statistical parameter that the smaller it gets the better the final model. Fig. 1 illustrates the computed MSE for the randomly selected training and validation data with respect to the tested tuning parameters for the deployed DL. It seems that the following tuning parameters give the most accepted results: lasso parameter ($\ell 1$) = 10^{-5}, activation = "rectifier with dropout", hidden = 100, input_dropout_ratio = 0.05, hidden_dropout_ratios = 0.4. In addition, cross-validation was performed; cross-validation is a method that divides the validation data in a number of groups (in this study, the number of groups (=folds) was chosen and kept to be 6). For

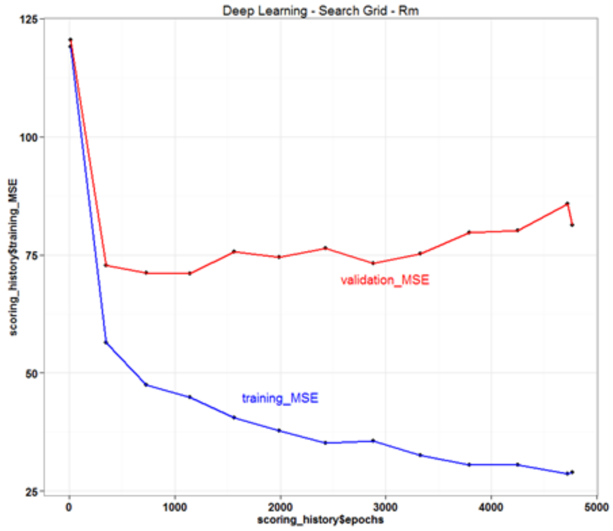

Fig. 1. Mean-squared error (MSE) for the tensile-strength (Rm) training and validation data as derived from the tuning of the deep learning (DL) model.

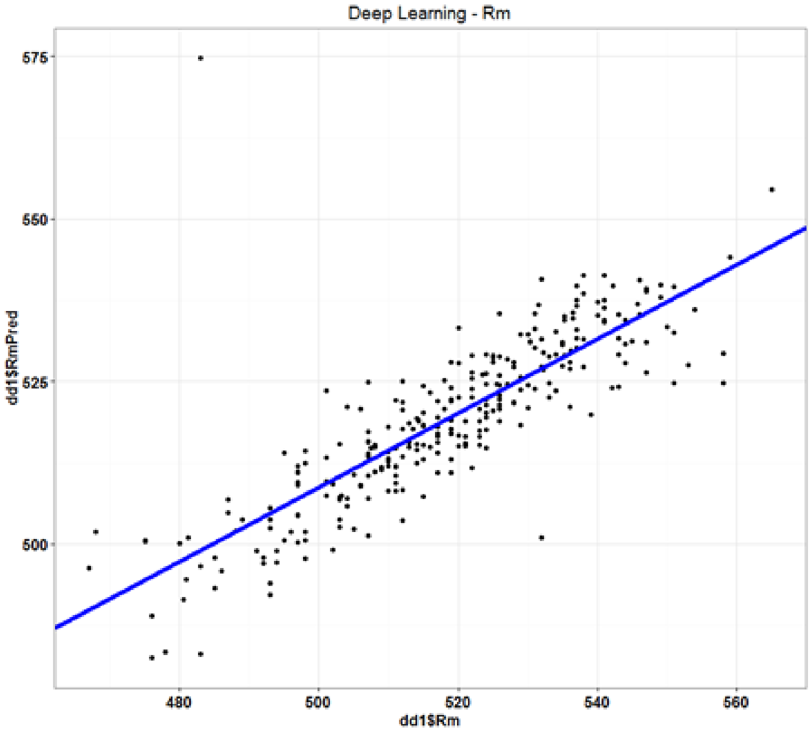

Fig. 2. Actual and predicted tensile-strength (Rm) values according to the deduced supervised deep learning (DL) model.

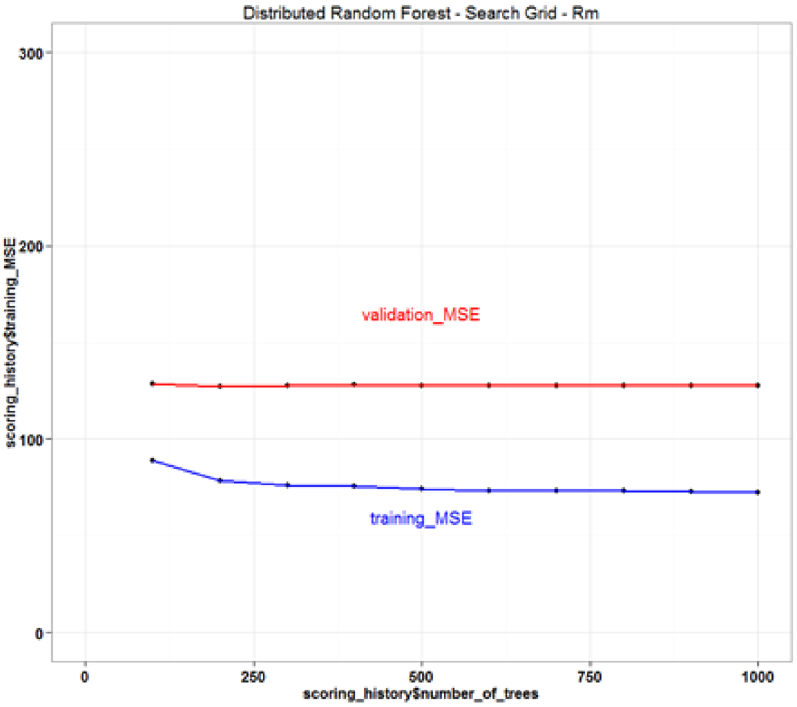

Fig. 3. Mean-squared error (MSE) for the tensile-strength (Rm) training and validation data as derived from the tuning of the distributed random forest (DRF) model.

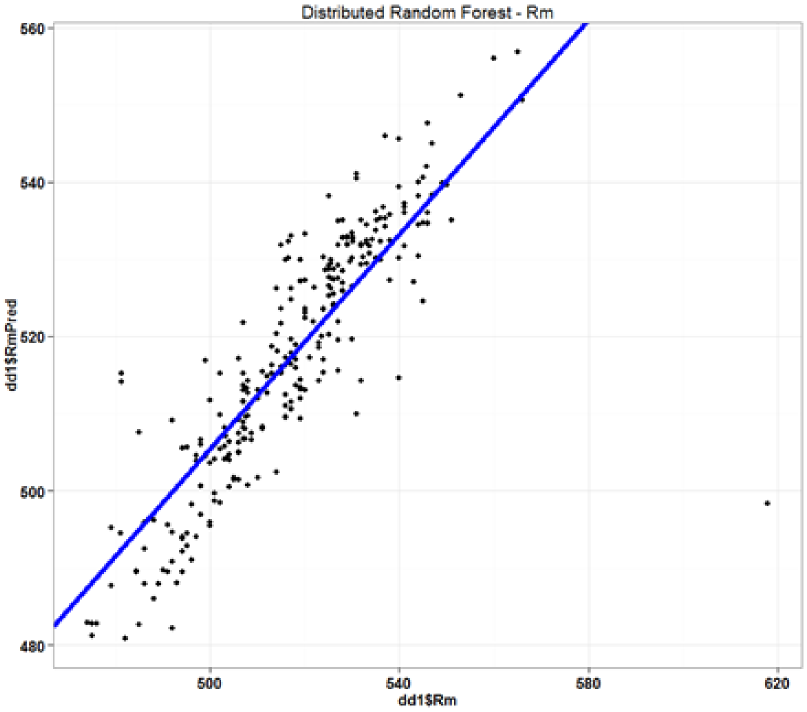

Fig. 4. Actual and predicted tensile-strength (Rm) values according to the deduced supervised distributed random forest (DRF) model.

the final accepted model the values of R2 = 0.91, and MSE = 30.3 were obtained from the training data; for the validation data the corresponding values were 0.66, and 118.7. Consequently, some overfitting of the data takes place. This happens when the model picks up some noise from the training data and sticks with that. Fig. 2 depicts the actual and predicted Rm values with respect to the finally accepted model.

The tuning parameters for the DRF that seemed to be the best were: sample_rate = 0.96, min_rows = 4, ntrees = 1000, max_depth = 24, and col_sample_rate_per_tree = 1. Fig. 3 depicts the training and validation values of MSE from the searching runs. The selected model gave the values of R2 = 0.73, and MSE = 92.7 for the training data, and from the validation data the values of R2 = 0.70, and MSE = 107.8 were deduced. Fig. 4 illustrates the actual and predicted Rm values based on the selected final model.

For the GBM the tuning parameters that showed the best overall results were: sample_rate = 0.5, min_rows = 10, ntrees = 300, max_depth = 20, learn_rate = 0.1, col_sample_rate = 0.5, and col_sample_rate_per_tree = 0.3. Fig. 5 presents the training and validation MSE values in the search process for the tuning parameters. The final accepted model gave R2 = 0.87, MSE = 44.8, for the training data, and R2 = 0.81, MSE = 61.2, for the validation data. The actual and predicted Rm values by the model are shown in Fig. 6.

Yield stress (Re) results

The DL tuning parameters that appeared to give the best model were: lasso parameter (ℓ1) = 10^{-5}, activation = "rectifier with dropout", hidden = 150, input_dropout_ratio = 0.10, hidden_dropout_ratios = 0.3. Fig. 7 presents the training and validation MSE values as deduced during the fine tuning search process. The final accepted model gave R2 = 0.92, MSE = 51.1 for the training data, and R2 = 0.66, and MSE = 200.7 for the validation data. Overfitting seemed to be unavoidable in the deployment of the DL model. Fig. 8 illustrates the actual and predicted Re values according to the final selected DL model.

The DRF tuning parameters that seemed to give good results were: sample_rate = 0.96, min_rows = 4, ntrees= 1000, max_depth = 27, col_sample_rate_per_tree = 0.9. Fig. 9 shows the training and validation MSE values obtained during the tuning process. The final accepted model gave the values of R2 = 0.66, MSE = 209.8,

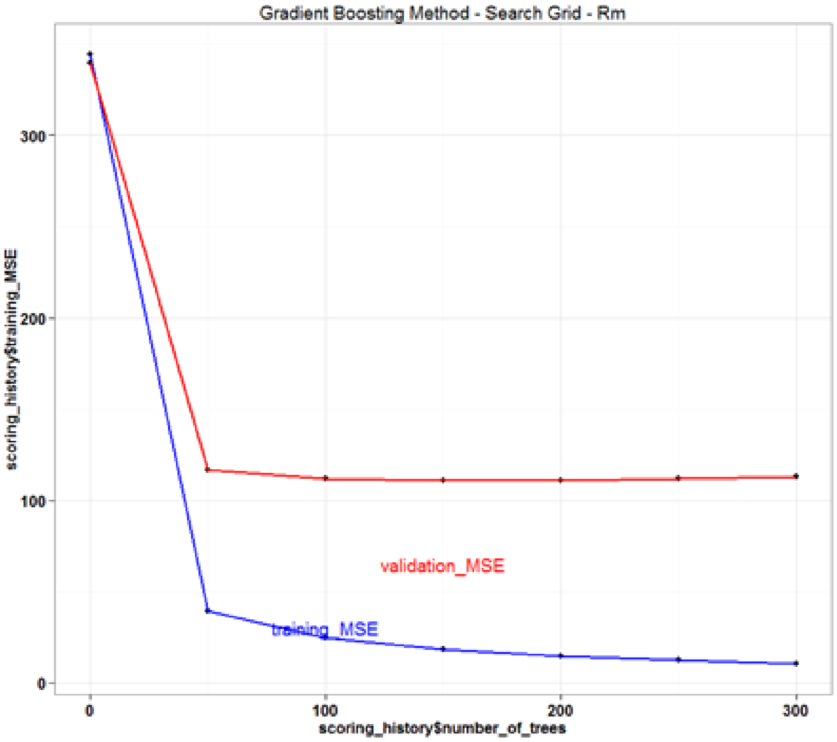

Fig. 5. Mean-squared error (MSE) for the tensile-strength (Rm) training and validation data as derived from the tuning of the gradient boosting method (GBM) model.

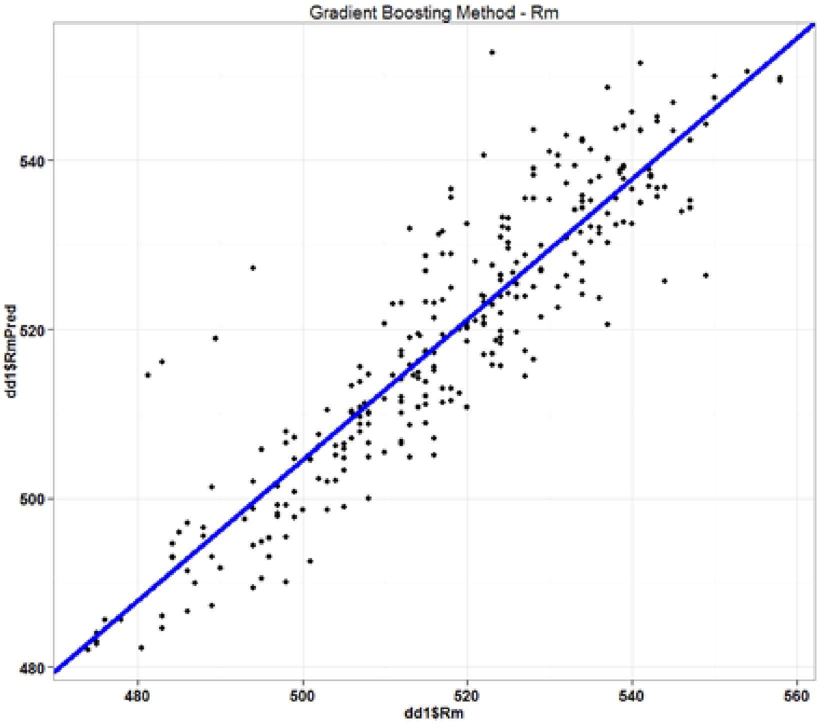

Fig. 6. Actual and predicted tensile-strength (Rm) values according to the deduced supervised gradient boosting method (GBM) model.

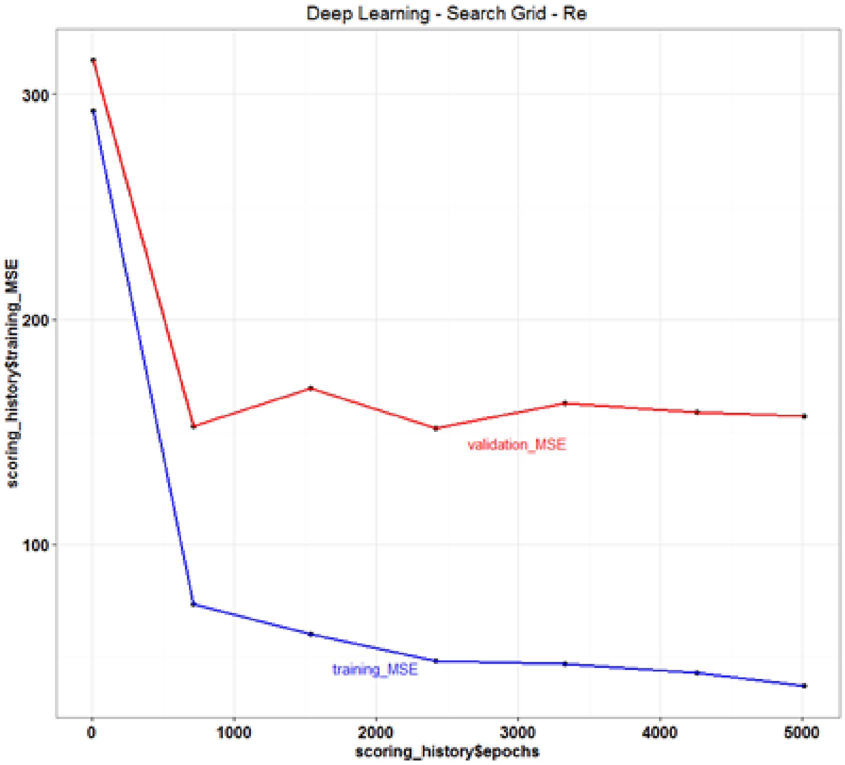

Fig. 7. Mean-squared error (MSE) for the yield-stress (Re) training and validation data as derived from the tuning of the deep learning (DL) model.

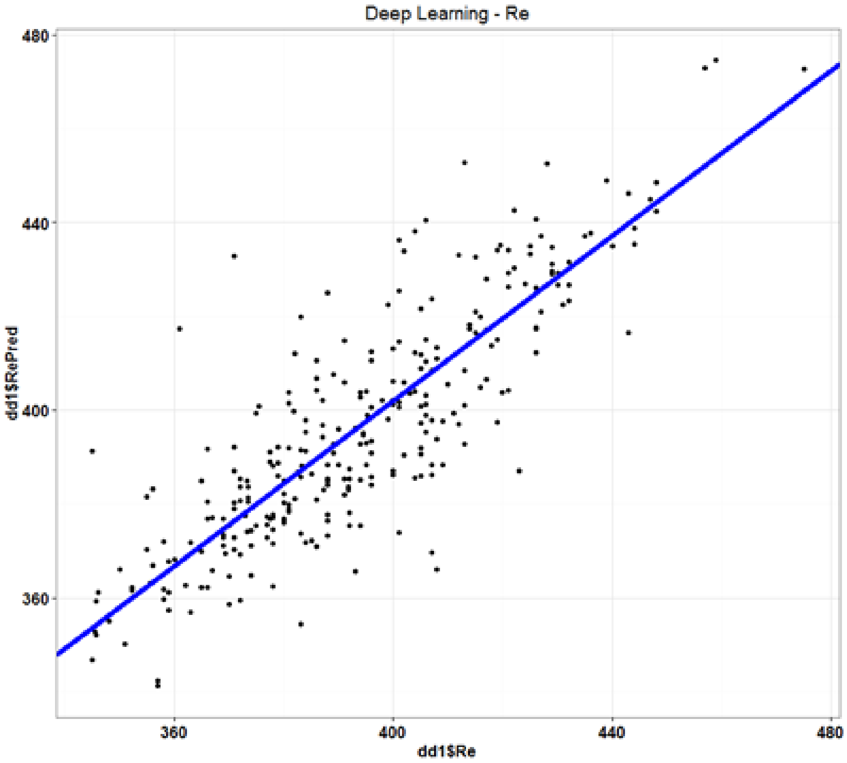

Fig. 8. Actual and predicted yield-stress (Re) values according to the deduced supervised deep learning (DL) model.

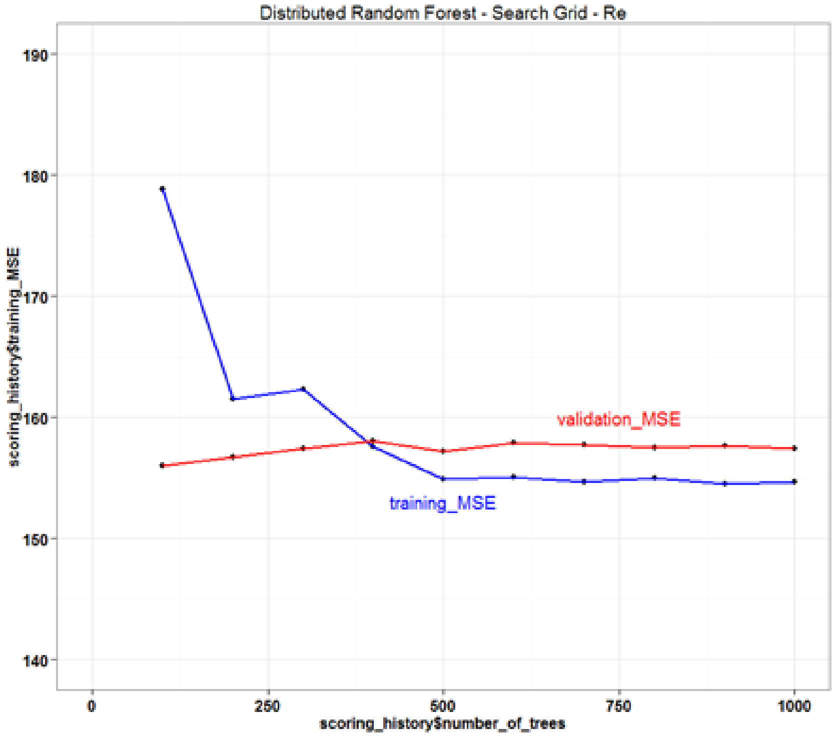

Fig. 9. Mean-squared error (MSE) for the yield-stress (Re) training and validation data as derived from the tuning of the distributed random forest (DRF) model.

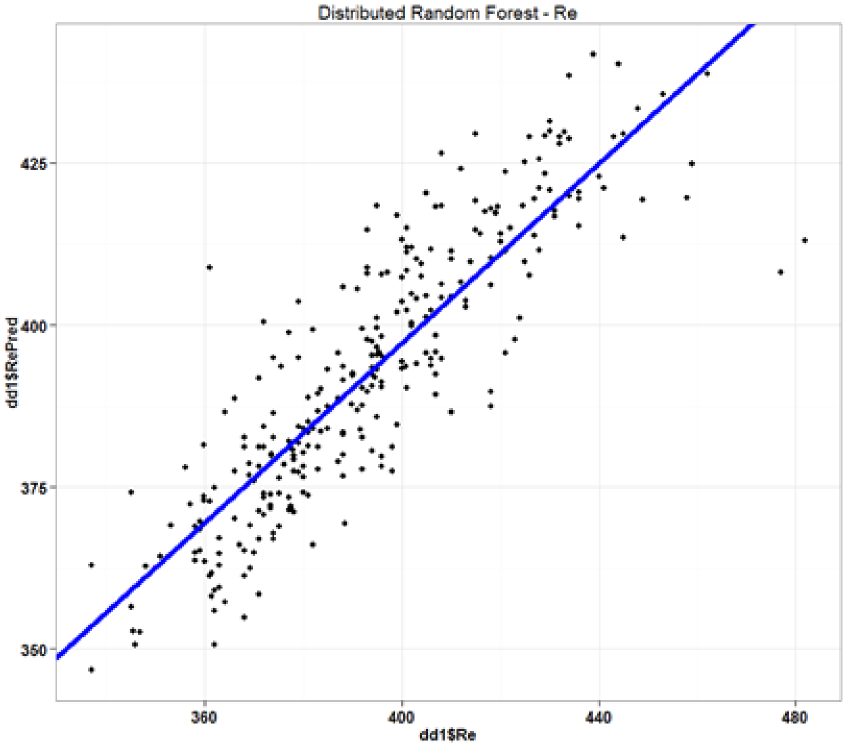

Fig. 10. Actual and predicted yield-stress (Re) values according to the deduced supervised distributed random forest (DRF) model.

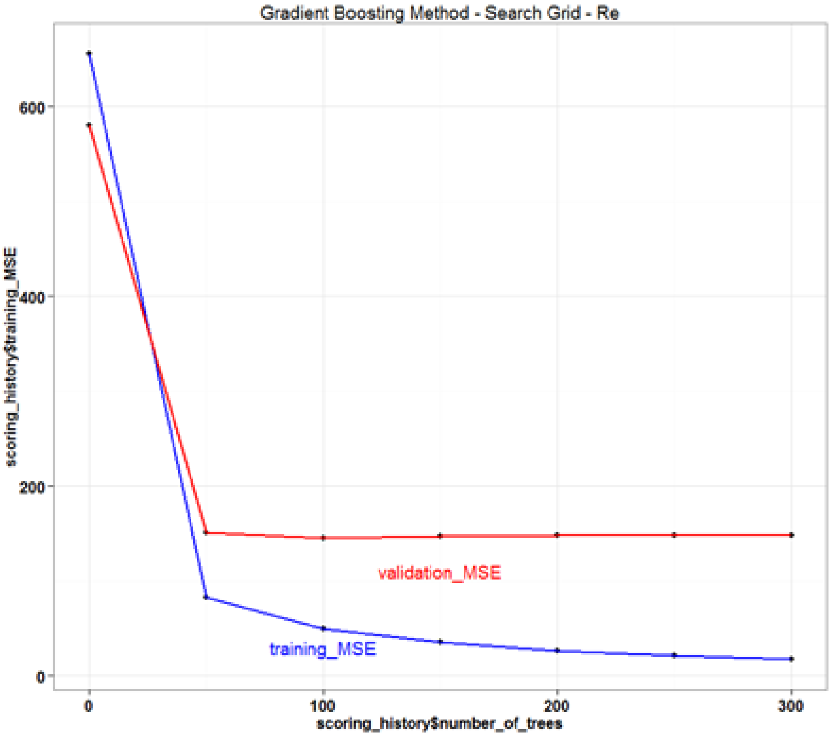

Fig. 11. Mean-squared error (MSE) for the yield-stress (Re) training and validation data as derived from the tuning of the gradient boosting method (GBM) model.

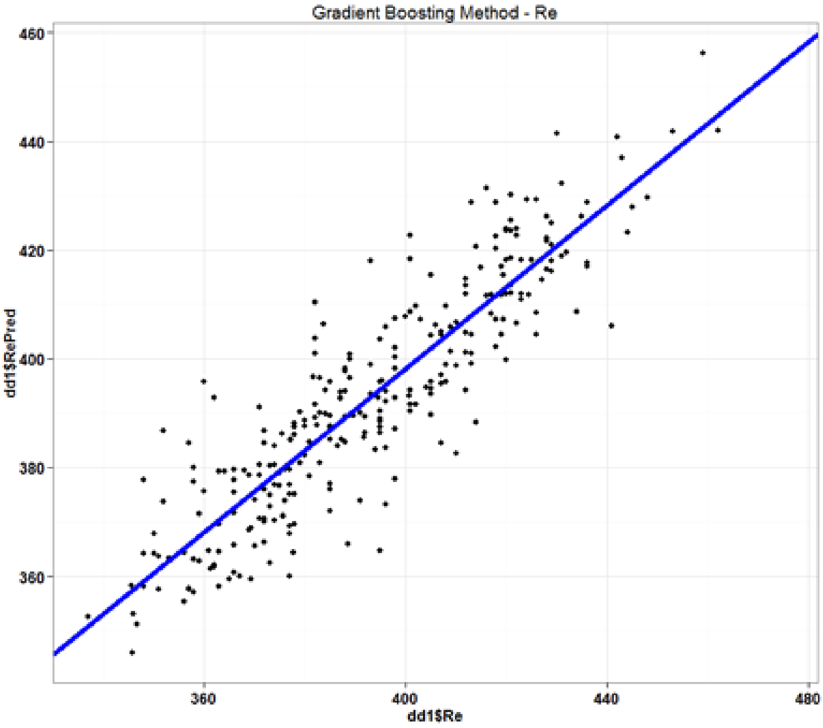

Fig. 12. Actual and predicted yield-stress (Re) values according to the deduced supervised gradient boosting method (GBM) model.

Table 1. The 10 most important parameters for Rm as derived by the models.

	Grade S355: Tensile strength (Rm)		
No.	DL	DRF	GBM
1	TARGET_WIDTH	Thickness	WAITING_THICK_1
2	ROLLING_MODE.MAN	Ceq	TARGET_LENGTH_L3
3	Grade_S355.S355J2C+N	WAITING_THICK_1	Ceq
4	Al	TARGET_LENGTH_L3	Thickness
5	Thickness	V	ENTRY_WIDTH
6	FINISHING_TEMP_PASS	C	C
7	ENTRY_WIDTH	Nb	V
8	HEATING_TIME	REASTART_TEMP_1	REASTART_TEMP_1
9	PASS_NUMBER	ENTRY_WIDTH	As
10	TARGET_LENGTH_L3	Cu	ActiveO2ppm

for the training data, and R2 = 0.75, MSE = 170.7, for the validation data. Actual and predicted Re values by the finally selected DRF model are presented in Fig. 10.

Finally, the tuning parameters for the GBM model that showed the best overall results were: sample_rate = 0.5, min_rows = 10, ntrees = 300, max_depth = 10, learn_rate = 0.1, col_sample_rate = 0.3, col_sample_rate_per_tree = 0.5. Fig. 11 illustrates the training and validation MSE values during the tuning selection process. The final accepted GBM model resulted into the following values: R2 = 0.89, MSE = 71.0, for the training data, and R2 = 0.80, MSE = 130.7, for the validation data. Fig. 12 depicts the actual and predicted Re values with respect to the deduced GBM supervised model.

Variable importances

The term "variable importances" is mostly technical, and for this reason the 's' is kept, although grammatically seems not to be correct. In any case, what is important is the list of the most important variables, or in other words, the technological parameters that seem to affect the examined mechanical properties Rm, and Re. As stated above, 33 independent parameters from the meltshop and the plate mill were selected in order to be included in this statistical analysis. However, only the top 10 from the list were selected to be presented in this study. Tables 1 and 2 present these results for the Rm and Re properties. What is mostly intriguing is that some technological parameters appear in all 3 selected algorithms

Table 2. The 10 most important parameters for Re as derived by the models.

	Grade S355: Yield stress (Re)		
No.	DL	DRF	GBM
1	ROLLING_MODE.MAN	Thickness	WAITING_THICK_1
2	PRODUCT_TYPE.DEFAULT	TARGET_LENGTH_L3	TARGET_LENGTH_L3
3	Thickness	WAITING_THICK_1	Thickness
4	ENTRY_WIDTH	Nb	TARGET_WIDTH
5	Grade_S355.S355J2C+N	**C**	Ceq
6	TARGET_WIDTH	V	ENTRY_WIDTH
7	TARGET_LENGTH_L3	Ceq	C
8	PASS_NUMBER	ENTRY_WIDTH	V
9	WAITING_THICK_1	REASTART_TEMP_1	Cr
10	SPH	ActiveO2ppm	Mn

per specific mechanical property; these are presented with bold letters. In bold italics are presented parameters that appear in the two out of the three deployed models. There are parameters, for example 'Thickness', that are expected to be in the list. Practice has shown that the mechanical properties tend to increase by decreasing sizes. However, there are parameters that seem to be important and their role is not so obvious from the first point of view or the experience gathered throughout the years. One such parameter is the 'ActiveO2ppm' that reflect to the EAF ppmO levels at tapping. Although ferroalloys are added at tapping to deoxidize liquid steel, and active ppm-oxygen values actually drop below 10, it rather suggests the effect from the steel cleanliness related to carbon levels at scrap meltdown. This is something that is being recognized more and more by time worldwide, lately. Continuing discussion, Fig. 13 presents in graphical form the relationship between the two important parameters "Thickness", and "REASTART_TEMP_1". A 3rd-degree spline curve is shown in order to show the potential correlation. Only the salient features are presented, that is why it is not a solid line. The plate-mill automation parameter "REASTART_TEMP_1" actually means the temperature at which the thermomechanical rolling re-starts. It is obvious that by increasing the final plate thickness one has to wait a bit more for the temperature to drop at lower values. Another very interesting aspect between "Thickness" and "TAR-

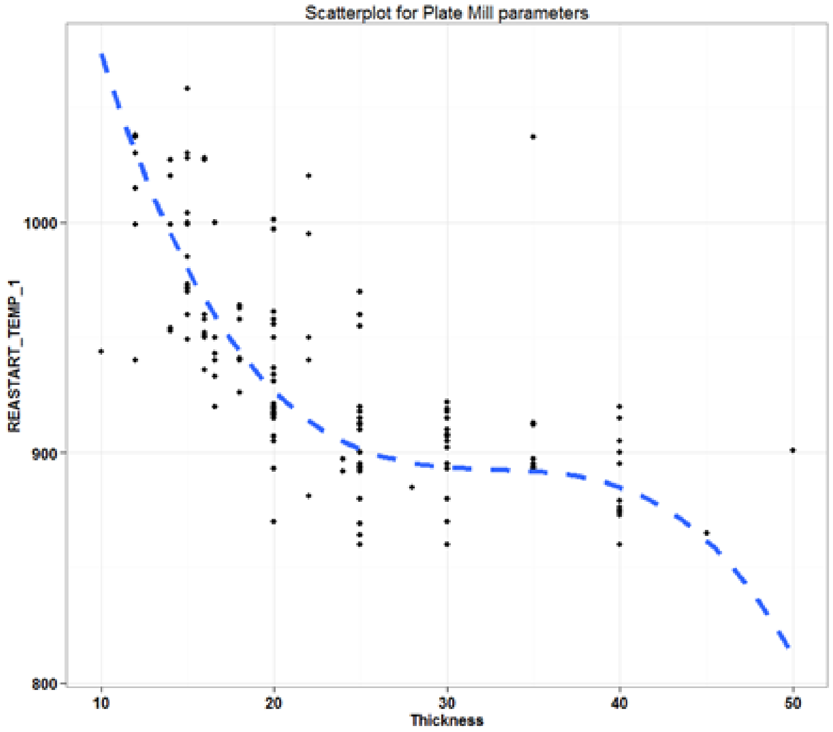

Fig. 13. Scatter-plot of the plate technological parameters under rolling: “Thickness” versus “REASTART_TEMP_1” parameters.

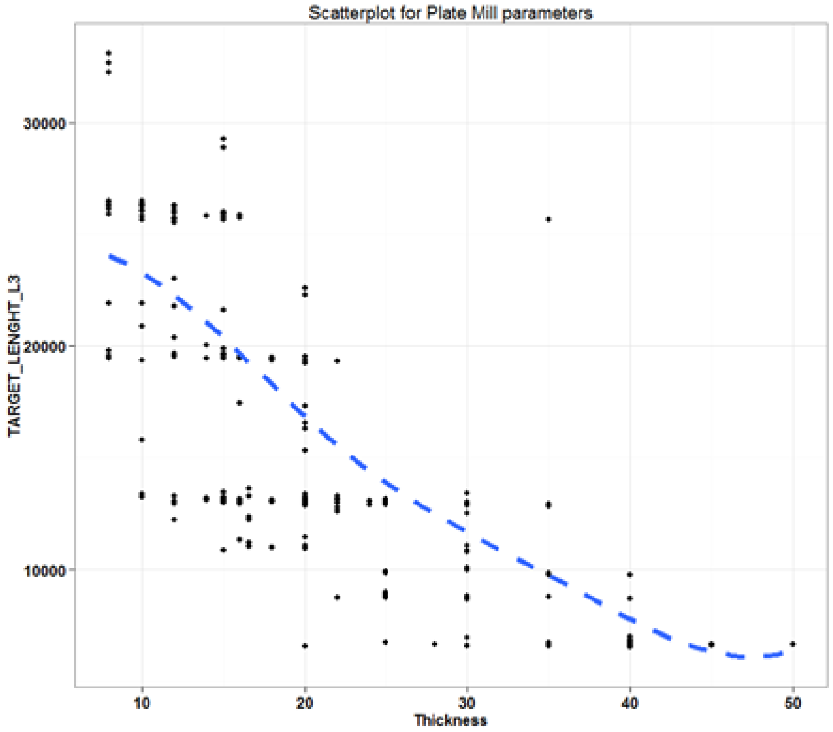

Fig. 14. Scatter-plot of the plate technological parameters under rolling: “Thickness” versus “TARGET_LENGTH_L3” parameters.

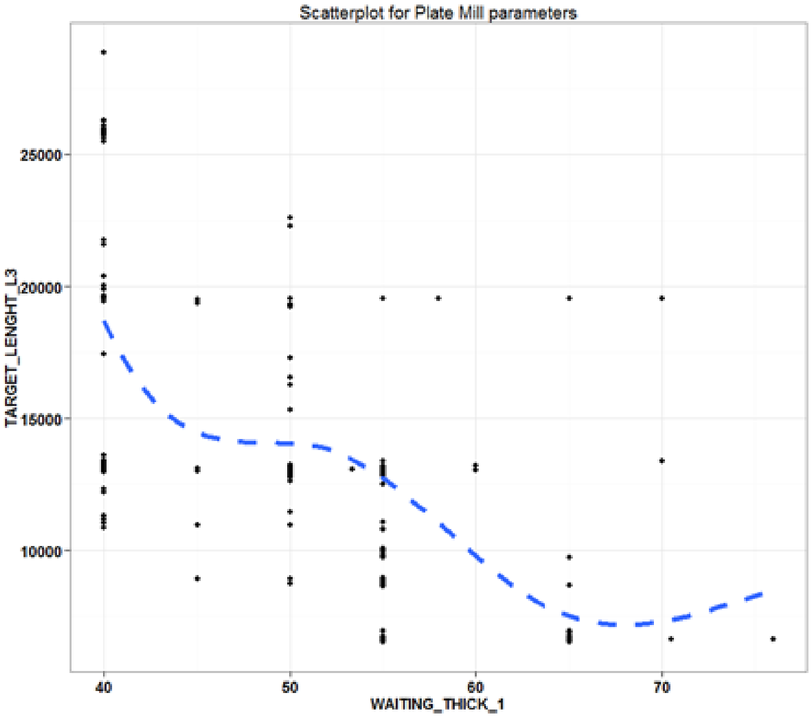

Fig. 15. Scatter-plot of the plate technological parameters under rolling: “WAITING_THICK_1” versus “TARGET_LENGTH_L3” parameters.

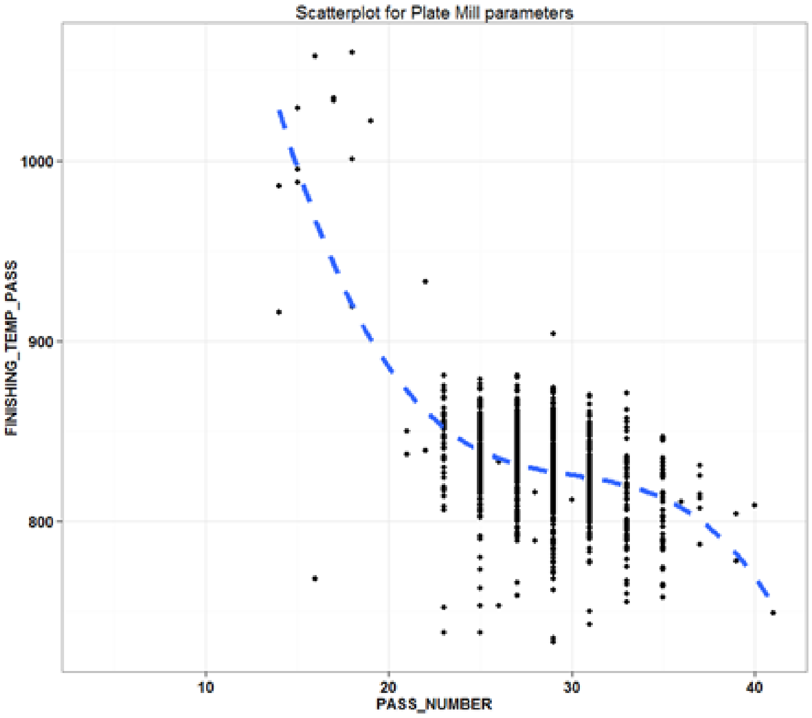

Fig. 16. Scatter-plot of the plate technological parameters under rolling: “PASS_NUMBER” versus “FINISHING_TEMP_PASS” parameters.

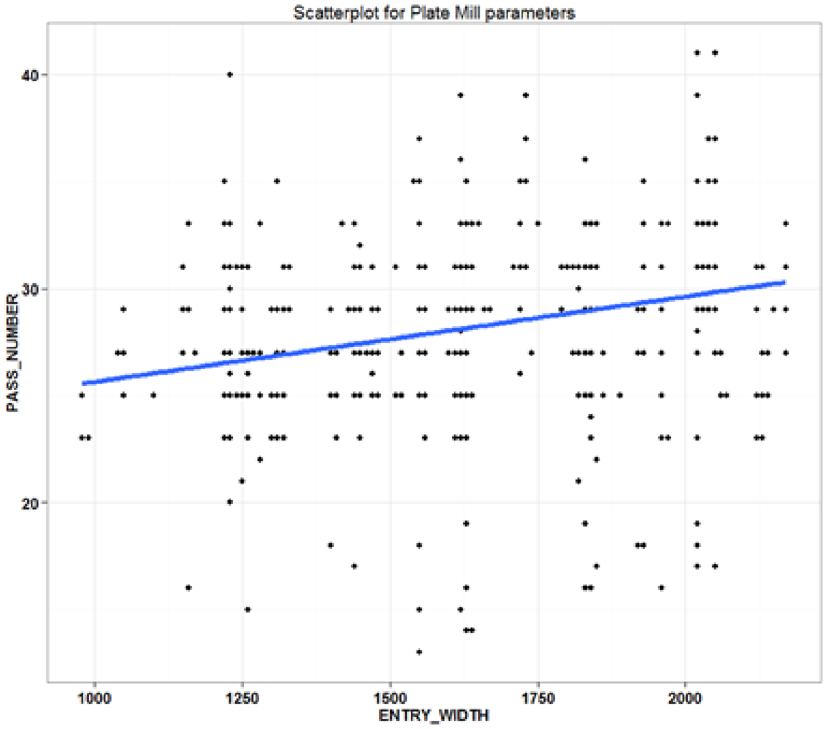

Fig. 17. Scatter-plot of the plate technological parameters under rolling: “ENTRY_WIDTH” versus “PASS_NUMBER” parameters.

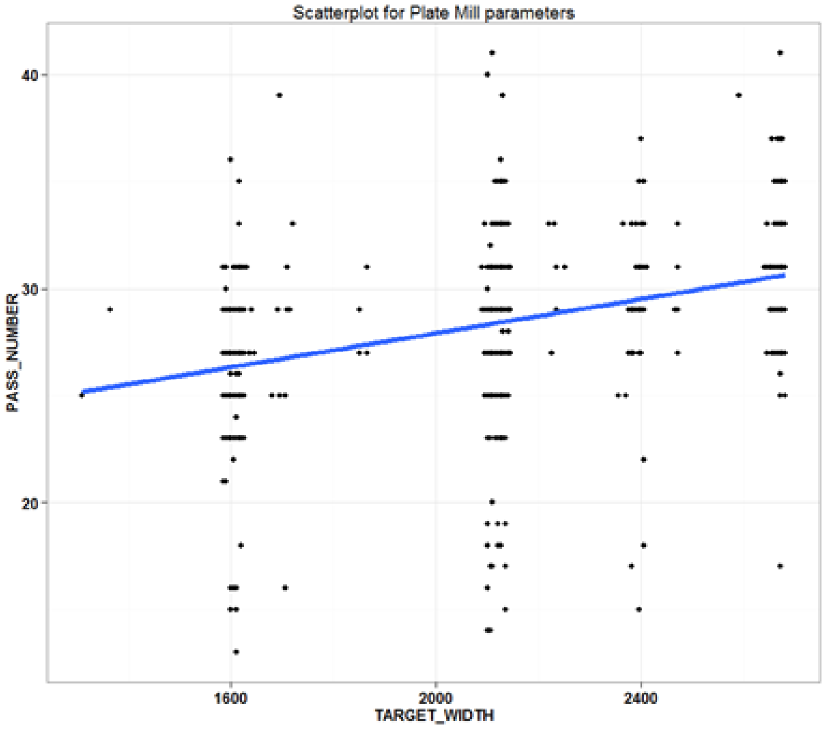

Fig. 18. Scatter-plot of the plate technological parameters under rolling: “TARGET_WIDTH” versus “PASS_NUMBER” parameters.

GET_LENGTH_L3" is depicted in Fig. 14. A 3rd-degree spline curve is shown in order to show just the salient features. It seems that larger target lengths were produced for smaller thickness plates and this trend most probably drew the "TARGET_LENGTH_L3" parameter into the models. Apart from the restarting temperature in thermomechanical rolling (TMR), the intermediate plate thickness at which the rolling pauses for cooling is critical, as well; this is the "WAITING_THICK_1" and together with "TARGET_LENGTH_L3" are plotted in Fig. 15. It seems that some correlation may exist as of the 5th-degree spline curve also shown. Finding relations between these types of parameters is beyond imagination in real practice. One requires a robust statistical system to pin-point these subtle details. Fig. 16 shows the relation between "PASS_NUMBER" and "FINISHING_TEMP_PASS" which are two more important variables; a 3rd-degree spline curve is presented, as well. This is somewhat expected as the larger the number of passes the smaller the plate temperature is expected to be. Two very intriguing relationships have been found between "PASS_NUMBER" and "ENTRY_WIDTH", and "TARGET_WIDTH" parameters. These relationships are pretty subtle yet the straight lines as presented in Fig. 17, and 18, were deduced from strong statistical correlations. In fact, an analysis of variance (ANOVA) [7] for the straight line correlation presented Fig. 17 gave a standard error of 3.5 with an F-distribution value $F_{1,1139} = 123.9$ ($p < 2.2\ 10^{-16}$). Similarly, an ANOVA for the correlation presented in Fig. 18 gave a standard error of 3.4 with an F-distribution value $F_{1,1139} = 194$ ($p < 2.2\ 10^{-16}$). After this type of analysis it may be realized why "ENTRY_WIDTH", and "TARGET_WIDTH" enter into the picture of important predictors. They are both being dragged by the important parameter "PASS_NUMBER". In this study, the GBM appears to have generated the most reliable models for the Rm and Re properties, as over-fitting was minimized in that case.

CONCLUSIONS

The mechanical properties of tensile strength and yield stress for our S355-based plate products together with 33 more technological parameters from the process involved were statistically analyzed with the help of the DL, DRF, and GBM algorithms supplied by the H2O Flow system. Models predicting the mechanical properties under investigation were derived together with a list of the ten most important technological parameters that seem to affect them. Some parameters,not considered in advance, seem to play an important role in the process and should be followed up in the future campaigns to verify their effect in actual practice.

Acknowledgements

The support of the top management for this type of work is greatly appreciated.

REFERENCES

1. A. Candel, E. LeDell, V. Parmar, A. Arora, in: J. Lanford (Ed.), Deep Learning with H2O, published by H_2O.ai, Inc., Mountain View, California, 5th ed., 2016.
2. S. Aiello, E. Eckstrand, A. Fu, M. Landry, P. Aboyoun, in: J. Lanford (Ed.), Machine Learning with R and H2O, published by H2O.ai, Inc., Mountain View, California, 6th ed., 2016.
3. C. Click, M. Malohlava, A. Candel, H. Roark, V. Parmar, in J. Lanford (Ed.), Gradient Boosted Models with H_2O, published by H_2O.ai, Inc., Mountain View, California, 6th ed., 2016.
4. H_2O Flow version 0.4.26, H_2O Build project version 3.8.1.3, jenkins-rel-turan-3, www.h2o.ai.
5. J.F. Wiley, R Deep Learning Essentials, Birmingham-Mumbai, PACKT Publishing, UK, 2016.
6. N. Zumel, J. Mount, Practical Data Science with R, MANNING, Shelter Island, New York, 2014.
7. J. Faraway, Practical regression and ANOVA using R, University of Bath, UK, 2002.