Oct 28th, 8:00 AM - Oct 29th, 6:00 PM

Enhancing Real Estate Management: The Transformative Role of Machine Learning in Predictive Gains and Risk Model Performance

Visar Hoxha
University for Business and Technology - UBT, visar.hoxha@ubt-uni.net

Blerta Demjaha
blerta.demjaha@eukos.org

Veli Lecaj
University for Business and Technology - UBT, veli.lecaj@ubt-uni.net

Hazer Dana
University for Business and Technology, hazer.dana@ubt-uni.net

Fuat Pallaska
University for Business and Technology - UBT, fuat.pallaska@ubt-uni.net

Enhancing Real Estate Management: The Transformative Role of Machine Learning in Predictive Gains and Risk Model Performance

Visar Hoxha[1], Blerta Demjaha[2] Veli Lecaj[1] Hazer Dana[1] Fuat Pallaska[1]
[1] Faculty of Real Estate, UBT
Visar.hoxha@ubt-uni.net
[2] Real Estate Department, College ESLG
blerta.demjaha@eukos.org
[1] Faculty of Real Estate, UBT
Veli.lecaj@ubt-uni.net
[1] Faculty of Real Estate, UBT
hazer.dana@ubt-uni.net
[1] Faculty of Real Estate, UBT
fuat.pallaska@ubt-uni.net

Abstract: Mortgage scoring models are pivotal in evaluating the risk associated with mortgages. Traditionally, these models were constructed using logistic regression. However, with the rise of machine learning, algorithms such as classification trees and neural networks have been employed. These algorithms are trained on a sample of mortgages, with the occurrence or non-occurrence of default observed. The data is then split into training and test samples, with machine learning algorithms further dividing the training sample for validation. This approach aims to determine hyperparameters that maximize performance while minimizing overfitting. Once calibrated, the model is applied to the test sample to predict default events. Despite the sophistication of machine learning algorithms, their predictive performance in mortgage scoring is comparable to logistic regression. Ensemble methods, which combine multiple models, have shown potential in enhancing predictive performance. This literature review explores the application of machine learning in mortgage scoring, comparing it with traditional methods, and discussing its implications.

Keywords: scoring model, mortgage default, machine learning algorithms, logistic regression, receiver operating characteristic curve, neural networks

1. Introduction

The rapid evolution of machine learning and its application in various sectors has garnered significant attention in the academic and industrial world. One such application is in the realm of real estate mortgage scoring. Traditional scoring models, primarily based on logistic regression, have been the cornerstone for evaluating the risk associated with mortgages. However, with the advent of machine learning algorithms, there's a paradigm shift in how these evaluations are conducted. This literature review aims to delve into the nuances of machine learning algorithms in mortgage scoring, comparing their efficacy with traditional models, and understanding the intricacies of their predictive performance.

2. Literature review

Any scoring model is constructed from a sample of n mortgages for which the occurrence or non-occurrence of default is observed, represented by a dichotomous variable Y. For each individual in this sample, we also have a set of explanatory or predictor variables, which correspond, for example, to information on the nature of the contract and the borrower. This database is then broken down into two sub-samples: a training sample on which the model is selected, calibrated, and possibly estimated, and a test sample on which the out-of-sample predictive performance is evaluated (Foryś, 2022) [1]. For machine learning algorithms, the training sample is usually decomposed into two sub-samples: a sample on which the classification algorithm is trained and a validation sample that makes it possible to determine the value of the hyperparameters (or tuning parameters) associated with the classification method and thus to control the phenomenon of over-learning.

The idea is then to determine the value of the hyperparameters, which maximizes a performance measure calculated on a sample (the validation sample) different from that on which the algorithm is trained (the learning sample). Thus, this approach reduces the risk of overfitting induced by setting "optimal" values for the hyperparameters (Choy & Ho, 2023) [2]. This would allow the classification to be reproduced almost perfectly on the training sample but would ultimately lead to poor performance of out-of-sample classification. This approach can be generalized to a k-fold cross-validation approach applied to the entire learning sample.

Once the model has been calibrated (for machine learning algorithms) or estimated (for the usual parametric approaches), it is applied to the test sample. Depending on the models, we then obtain for each individual in the test sample either an estimate of the conditional probability of occurrence of the default event, as, for example, in the case of a logistic regression, or directly a forecast of this event represented in the form of a dichotomous variable $\hat{Y}$, as, for example, in the case of a classification tree. When the models produce estimated probabilities, we are reduced to a forecast on the event $\hat{Y}$ by comparing the probability to a threshold c, typically 50% (Foryś, 2022)[3]. If the probability exceeds this threshold, we predict the event's occurrence, i.e., $\hat{Y}(c) = 1$. For a given threshold, we can construct a confusion matrix listing the occurrences of two classification errors made on the test sample. False positives correspond to individuals

[1]Foryś, 2022
[2]Choy & Ho, 2023
[3]Foryś, 2022

for whom the model had predicted a defect ($\hat{Y}$ (c) = 1) but for whom no defect was observed ex-post (Y= 0).

Conversely, false negatives correspond to individuals for whom the model had not predicted a defect and for whom a defect was observed. These errors can be expressed as ratios, such as specificity and sensitivity. The sensitivity corresponds to the probability of predicting the defect in the population of defects. At the same time, the specificity is the probability of predicting a non-defect in the population of non-defects. From these elements, we can then construct the Receiver Operating Characteristic (ROC) curve, the elements of which correspond to the sensitivity (ordinate axis) and the specificity (abscissa axis) obtained for threshold values c varying from 0 to 1 (see graph below). The interest of the ROC curve is to allow the predictive capacity of the classification model to be assessed independently of the choice of the threshold.

From the mid-1980s, many academic studies sought to assess the predictive performance gains of machine learning methods compared to logistic regression. Thirty years later, the diagnosis is relatively mixed. Makowski (1985) [4], Coffman (1986) [5], Srinivasan and Kim (1987) [6], and Carter and Catlett (1987) [7] were among the first to apply classification trees for real estate mortgage scoring to capture the interactions between predictors (Foryś, 2022) [8]. Artificial neural networks were also very quickly applied, mainly to problems of scoring banking establishments (Tam and Kiang, 1992) [9] or companies.

This last study concludes in a mixed way by pointing in particular to the black box aspect of neural networks, the sometimes-illogical weight given to certain predictors, and the overfitting problems. Isada (2022) [10] compares different kinds of neural networks to standard techniques, such as logistic regression and linear discriminant analysis, on a personal real estate mortgage basis. They show that neural networks offer good predictive performance when looking at the percentage of correctly classified bad real estate mortgages. On the other hand, the predictive performance of neural networks is similar to that of logistic regression concerning the percentage of good and bad real estate mortgages correctly identified.

In general, individual machine learning classifiers do not significantly improve the predictive performance of logistic regression. These results are confirmed by Isada (2022) [11], who proposes the first literature synthesis concerning scoring models, including machine learning techniques. The author reports the Percentage of Correct Classification (PCC) of six methods (classification trees, neural networks, logistic regression, linear regression, etc.) from five studies. It shows that no method dominates the others, but the differences between the PCCs of these different methods are very small.

[4] Makowski, 1985
[5] Coffman, 1986
[6] Srinivasan and Kim, 1987
[7] Carter and Catlett, 1987
[8] Foryś, 2022
[9] Tam and Kiang, 1992
[10] Isada, 2022
[11] Ibid

These results are confirmed by the comparative study by Baesens et al.(2003) [12], which offers a systematic analysis of seventeen classification algorithms from eight mortgage databases provided by international banks. Support Vector Machines (SVM) or neural networks offer very good predictive performance for most of the databases considered, with Area under the Curve (AUC) ranging from 66% to 91% (Koktashev et al., 2019) [13]. But the authors also show that the differences between the AUC of the best machine learning method and that of the logistic regression are less than 2% for most bases.

How can we explain such low-contrast predictive performance? The main advantage of these machine learning algorithms over standard parametric approaches lies in their ability to automatically reveal interactions between predictors and nonlinearities (threshold effects). Consider the example of a classification tree such as the one shown in the diagram above. Classifying a real estate mortgage as bad or good risk takes the form of a tree that splits into two at each node. The value of a predictor (for example, residential status) determines whether the right branch (non-owner) or the left branch (owner) should be considered for the rest of the algorithm. At the end of the algorithm, when the last node is reached, the real estate mortgage is assigned to a leaf and a forecast (0 or 1) (Koktashev et al., 2019) [14]. This forecast corresponds to the majority class (0 or 1) of the observations belonging to this node. For example, imagine that out of the 1,000 mortgages in the initial sample, 120 mortgages were granted to customers as follows:

(1) owners
(2) customers with more than two years of seniority in the bank, and
(3) customers without children.

If among those 120 real estate mortgages assigned to the left leaf of the tree, the default frequency is low, for example, 14%, then the absence of default is predicted for all the real estate mortgages having these characteristics. Ultimately, everything happens as if we were considering a regression model in which binary explanatory variables defined by-products (or interactions) of the initial predictors would be introduced (Koktashev et al., 2019) [15]. For example, the schema tree above ultimately amounts to constructing a first explanatory variable equal to 1 if the customer is the owner, has more than two years of seniority in his bank, and has no children. Thus, the classification trees make it possible to capture interactions between the initial predictors and nonlinear effects, typically threshold effects in this case, which would have been difficult to identify in a standard parametric approach without evaluating an infinitely large number of combinations and thresholds. In general, we find a similar idea in many machine learning algorithms (neural networks, support vector machines, etc.) through transforming the data representation space.

The question is then to know if the modeled event presents this type of non-linearity is that real estate mortgage scoring is a field of application in which there are ultimately too few non-

[12] Baesens et al., 2003
[13] Koktashev et al., 2019
[14] Ibid
[15] Ibid

linearities in the usual data for the predictive performance gains of machine learning to be significant (Milunovich, 2019) [16].

Ultimately, the use of the first ensemble methods in the 2000s made it possible to obtain significant predictive gains. The intuition of these approaches is to combine different elementary classification models likely to provide additional information. We thus find the idea of an automatic combination of forecasts or models. Twelve years after the study by Baesens et al. (2003) [17], Lessman et al. (2015) [18] propose a new comparative analysis using other evaluation criteria (Brier score, H-measure, etc.) and the most recent machine learning algorithms, including ensemble methods based on the principle of bagging or boosting. In the end, their study focuses on 41 classification algorithms applied to 8 databases of real estate mortgages to individuals. Their conclusion favors machine learning is more than machine learning: several ensemble methods predict risk significantly better than logistic regression. For example, random forests systematically dominate individual classifiers, whether the latter are parametric (logistic regression) or of machine learning type [trees, neural networks, Support Vector Machines (SVM), etc.]. The best performance is obtained for heterogeneous ensemble methods like the Weighted Average Ensemble method. The second lesson of this study is that the gains in predictive performance linked to machine learning tend to level off. Methodological refinements of machine learning algorithms do not necessarily improve the performance of scoring models. For example, the Area Under Curves (AUCs) of rotation forests do not differ significantly from those of random forests.

The central question remains why do some machine learning algorithms exhibit good predictive performance? The answer is not obvious, and no rule seems to emerge. To date, no research has been able to explain the performance of these classifiers according to their characteristics and the characteristics of the databases.

Any scoring model is constructed from a sample of n mortgages for which the occurrence or non-occurrence of default is observed, represented by a dichotomous variable Y. For each individual in this sample, we also have a set of explanatory or predictor variables, which correspond, for example, to information on the nature of the contract and the borrower. This database is then broken down into two sub-samples: a training sample on which the model is selected, calibrated, and possibly estimated, and a test sample on which the out-of-sample predictive performance is evaluated (Foryś, 2022) [19]. For machine learning algorithms, the training sample is usually decomposed into two sub-samples: a sample on which the classification algorithm is trained and a validation sample that makes it possible to determine the value of the hyperparameters (or tuning parameters) associated with the classification method and thus to control the phenomenon of over-learning.

The idea is then to determine the value of the hyperparameters, which maximizes a performance measure calculated on a sample (the validation sample) different from that on which the

[16] Milunovich, 2019
[17] Baesens et al., 2003
[18] Lessman et al., 2015
[19] Foryś, 2022

algorithm is trained (the learning sample). Thus, this approach reduces the risk of overfitting induced by setting "optimal" values for the hyperparameters (Choy & Ho, 2023) [20]. This would allow the classification to be reproduced almost perfectly on the training sample but would ultimately lead to poor performance of out-of-sample classification. This approach can be generalized to a k-fold cross-validation approach applied to the entire learning sample.

Once the model has been calibrated (for machine learning algorithms) or estimated (for the usual parametric approaches), it is applied to the test sample. Depending on the models, we then obtain for each individual in the test sample either an estimate of the conditional probability of occurrence of the default event, as, for example, in the case of a logistic regression, or directly a forecast of this event represented in the form of a dichotomous variable $\hat{Y}$, as, for example, in the case of a classification tree. When the models produce estimated probabilities, we are reduced to a forecast on the event $\hat{Y}$ by comparing the probability to a threshold c, typically 50% (Foryś, 2022)[21]. If the probability exceeds this threshold, we predict the event's occurrence, i.e., $\hat{Y}(c) = 1$. For a given threshold, we can construct a confusion matrix listing the occurrences of two classification errors made on the test sample. False positives correspond to individuals for whom the model had predicted a defect ($\hat{Y}(c) = 1$) but for whom no defect was observed ex-post ($Y= 0$).

Conversely, false negatives correspond to individuals for whom the model had not predicted a defect and for whom a defect was observed. These errors can be expressed as ratios, such as specificity and sensitivity. The sensitivity corresponds to the probability of predicting the defect in the population of defects. At the same time, the specificity is the probability of predicting a non-defect in the population of non-defects. From these elements, we can then construct the Receiver Operating Characteristic (ROC) curve, the elements of which correspond to the sensitivity (ordinate axis) and the specificity (abscissa axis) obtained for threshold values c varying from 0 to 1 (see graph below). The interest of the ROC curve is to allow the predictive capacity of the classification model to be assessed independently of the choice of the threshold.

From the mid-1980s, many academic studies sought to assess the predictive performance gains of machine learning methods compared to logistic regression. Thirty years later, the diagnosis is relatively mixed. Makowski (1985) [22], Coffman (1986) [23], Srinivasan and Kim (1987) [24], and Carter and Catlett (1987) [25] were among the first to apply classification trees for real estate mortgage scoring to capture the interactions between predictors (Foryś, 2022) [26]. Artificial neural networks were also very quickly applied, mainly to problems of scoring banking establishments (Tam and Kiang, 1992) [27] or companies.

This last study concludes in a mixed way by pointing in particular to the black box aspect of neural networks, the sometimes-illogical weight given to certain predictors, and the overfitting

[20]Choy & Ho, 2023
[21]Foryś, 2022
[22]Makowski, 1985
[23]Coffman, 1986
[24]Srinivasan and Kim, 1987
[25]Carter and Catlett, 1987
[26]Foryś, 2022
[27]Tam and Kiang, 1992

problems. Isada (2022) [28] compares different kinds of neural networks to standard techniques, such as logistic regression and linear discriminant analysis, on a personal real estate mortgage basis. They show that neural networks offer good predictive performance when looking at the percentage of correctly classified bad real estate mortgages. On the other hand, the predictive performance of neural networks is similar to that of logistic regression concerning the percentage of good and bad real estate mortgages correctly identified.

In general, individual machine learning classifiers do not significantly improve the predictive performance of logistic regression. These results are confirmed by Isada (2022) [29], who proposes the first literature synthesis concerning scoring models, including machine learning techniques. The author reports the Percentage of Correct Classification (PCC) of six methods (classification trees, neural networks, logistic regression, linear regression, etc.) from five studies. It shows that no method dominates the others, but the differences between the PCCs of these different methods are very small.

These results are confirmed by the comparative study by Baesens et al.(2003) [30], which offers a systematic analysis of seventeen classification algorithms from eight mortgage databases provided by international banks. Support Vector Machines (SVM) or neural networks offer very good predictive performance for most of the databases considered, with Area under the Curve (AUC) ranging from 66% to 91% (Koktashev et al., 2019) [31]. But the authors also show that the differences between the AUC of the best machine learning method and that of the logistic regression are less than 2% for most bases.

How can we explain such low-contrast predictive performance? The main advantage of these machine learning algorithms over standard parametric approaches lies in their ability to automatically reveal interactions between predictors and nonlinearities (threshold effects). Consider the example of a classification tree such as the one shown in the diagram above. Classifying a real estate mortgage as bad or good risk takes the form of a tree that splits into two at each node. The value of a predictor (for example, residential status) determines whether the right branch (non-owner) or the left branch (owner) should be considered for the rest of the algorithm. At the end of the algorithm, when the last node is reached, the real estate mortgage is assigned to a leaf and a forecast (0 or 1) (Koktashev et al., 2019) [32]. This forecast corresponds to the majority class (0 or 1) of the observations belonging to this node. For example, imagine that out of the 1,000 mortgages in the initial sample, 120 mortgages were granted to customers as follows:

(1) owners
(2) customers with more than two years of seniority in the bank, and
(3) customers without children.

[28] Isada, 2022
[29] Ibid
[30] Baesens et al., 2003
[31] Koktashev et al., 2019
[32] Ibid

If among those 120 real estate mortgages assigned to the left leaf of the tree, the default frequency is low, for example, 14%, then the absence of default is predicted for all the real estate mortgages having these characteristics. Ultimately, everything happens as if we were considering a regression model in which binary explanatory variables defined by-products (or interactions) of the initial predictors would be introduced (Koktashev et al., 2019) [33]. For example, the schema tree above ultimately amounts to constructing a first explanatory variable equal to 1 if the customer is the owner, has more than two years of seniority in his bank, and has no children. Thus, the classification trees make it possible to capture interactions between the initial predictors and nonlinear effects, typically threshold effects in this case, which would have been difficult to identify in a standard parametric approach without evaluating an infinitely large number of combinations and thresholds. In general, we find a similar idea in many machine learning algorithms (neural networks, support vector machines, etc.) through transforming the data representation space.

The question is then to know if the modeled event presents this type of non-linearity is that real estate mortgage scoring is a field of application in which there are ultimately too few non-linearities in the usual data for the predictive performance gains of machine learning to be significant (Milunovich, 2019) [34].

Ultimately, the use of the first ensemble methods in the 2000s made it possible to obtain significant predictive gains. The intuition of these approaches is to combine different elementary classification models likely to provide additional information. We thus find the idea of an automatic combination of forecasts or models. Twelve years after the study by Baesens et al. (2003) [35], Lessman et al. (2015) [36] propose a new comparative analysis using other evaluation criteria (Brier score, H-measure, etc.) and the most recent machine learning algorithms, including ensemble methods based on the principle of bagging or boosting. In the end, their study focuses on 41 classification algorithms applied to 8 databases of real estate mortgages to individuals. Their conclusion favors machine learning is more than machine learning: several ensemble methods predict risk significantly better than logistic regression. For example, random forests systematically dominate individual classifiers, whether the latter are parametric (logistic regression) or of machine learning type [trees, neural networks, Support Vector Machines (SVM), etc.]. The best performance is obtained for heterogeneous ensemble methods like the Weighted Average Ensemble method. The second lesson of this study is that the gains in predictive performance linked to machine learning tend to level off. Methodological refinements of machine learning algorithms do not necessarily improve the performance of scoring models. For example, the Area Under Curves (AUCs) of rotation forests do not differ significantly from those of random forests.

The central question remains why do some machine learning algorithms exhibit good predictive performance? The answer is not obvious, and no rule seems to emerge. To date, no research has been able to explain the performance of these classifiers according to their characteristics and the characteristics of the databases.

---

[33] Ibid
[34] Milunovich, 2019
[35] Baesens et al., 2003
[36] Lessman et al., 2015

3. Discussion

Machine learning algorithms, especially classification trees, neural networks, and support vector machines, have been lauded for their ability to automatically reveal interactions between predictors and capture nonlinearities. For instance, classification trees can intuitively segment data based on certain criteria, such as residential status or bank seniority, and make predictions based on these segments. Such an approach can capture interactions and threshold effects that might be overlooked in traditional parametric models.

However, the predictive performance of individual machine learning classifiers, when compared to logistic regression, doesn't show a significant improvement. Studies by Isada (2022) and Baesens et al. (2003) confirm this observation, with differences in predictive performance being marginal at best. The primary advantage of machine learning algorithms lies in their ability to detect interactions between predictors and nonlinearities, which might not be prevalent in real estate mortgage scoring data. This could explain the low-contrast predictive performance observed.

Interestingly, ensemble methods introduced in the 2000s have shown promise in enhancing predictive performance. By combining different elementary classification models, these methods aim to harness the collective power of multiple models, potentially offering more accurate and robust predictions.

4. Conclusion

While machine learning algorithms offer sophisticated tools for data analysis and prediction, their application in real estate mortgage scoring has shown mixed results. Individual classifiers, though adept at capturing nonlinearities and interactions, do not significantly outperform traditional logistic regression models in predictive performance. However, ensemble methods present a promising avenue for future research and application. As the field of machine learning continues to evolve, it's imperative to continually assess and adapt these algorithms to specific domains, ensuring that they provide tangible benefits over traditional methods.

References:

Baesens B., V an Gestel T., Viaene S., Stepanova M., Suykens J. and Vanthienen J. (2003), “Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring” , Journal of the Operational Research Society , vol. 54, no. 6, pp. 627-635.

Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. Land, 12(4), 740. https://doi.org/10.3390/land12040740

Foryś, I. (2022). Machine learning in house price analysis: regression models versus neural networks. Procedia Computer Science, 207(12), 435–445. https://doi.org/10.1016/j.procs.2022.09.078

Isada, F. (2022). The impact of inter-organisational network structures on research outcomes for artificial intelligence technologies. International Journal of Economic Sciences, 11(1), 1–18. https://ideas.repec.org/a/aop/jijoes/v11y2022i1p1-18.htmachine learning

Koktashev, V., Makeev, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019). Pricing modeling in the housing market with urban infrastructure effect. Journal of Physics: Conference Series, 1353(1), 012139.
https://doi.org/10.1088/1742-6596/1353/1/012139

Lessmann S., Baesens B., Seow H. V. and Thomas L. C. (2015), “Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: an Update of Research”, European Journal of Operational Research , vol. 247, No. 1, p. 124-136.

Milunovich, G. (2019). Forecasting Australian Real House Price Index: A Comparison of Time Series and Machine Learning Methods. SSRN Electronic Journal, 6(56). https://doi.org/10.2139/ssrn.3417527