MOODY'S ANALYTICS

ANALYSIS
04 SEPTEMBER, 2020

Prepared by

Olga Loiseau-Aslanidi
Olga.Loiseau-Aslanidi@moodys.com
Director

Natchie Subramaniam Thiagarajah
NatchieSubramaniam.Thiagarajah@moodys.com
Senior Economist

Vera Tolstova
Vera.Tolstova@moodys.com
Economist

Contact Us

Email
help@economy.com

U.S./Canada
+1.866.275.3266

EMEA
+44.20.7772.5454 (London)
+420.224.222.929 (Prague)

Asia/Pacific
+852.3551.3077

All Others
+1.610.235.5299

Web
www.economy.com
www.moodysanalytics.com

Automating Interpretable Machine Learning Scorecards

INTRODUCTION

Scorecard quality depends on not only model performance but also its interpretability. In this paper, we use our toolbox to build and compare the performance of four scorecard models. The benchmark model leverages a modified logistic regression with constraints imposed via supervised binning and variable selection. Three challenger models are built using decision tree, random forest and gradient boosting methods. We demonstrate that the interpretable benchmark model sacrifices little predictive power compared to the unconstrained challenger models. Meanwhile, the constraints are frequently violated by the challenger models, causing counterintuitive results for scorecards where the interpretation is critical.

Automating Interpretable Machine Learning Scorecards

BY OLGA LOISEAU-ASLANIDI, NATCHIE SUBRAMANIAM THIAGARAJAH, AND VERA TOLSTOVA

Scorecard quality depends on not only model performance but also its interpretability. In this paper, we use our toolbox to build and compare the performance of four scorecard models. The benchmark model leverages a modified logistic regression with constraints imposed via supervised binning and variable selection. Three challenger models are built using decision tree, random forest and gradient boosting methods. We demonstrate that the interpretable benchmark model sacrifices little predictive power compared to the unconstrained challenger models. Meanwhile, the constraints are frequently violated by the challenger models, causing counterintuitive results for scorecards where the interpretation is critical.

Introduction

Interpretation is a key requirement for robust and tractable scorecard models for risk management, regulatory compliance, strategy-setting, and product-marketing. Each characteristic included in the model must not only be a strong predictor that makes operational sense but also comply with *a priori* expectations or constraints.

Such constraints represent desirable patterns and relationships between the predictors and the score, based on business experience, industry trends and regulatory requirements. For instance, everything else being equal, higher-income customers are expected to have lower default probability; the default probability is typically higher amongst unemployed individuals; and lower credit quality is associated with higher frequency of late payment. Characteristics such as age, gender and country origin need to comply with the fair treatment principle, and monotonicity constraints can be used to achieve the desired pattern.

Nevertheless, the inclusion of various types of constraints is not a readily available option in rapidly evolving scorecard-building setups leveraging various machine learning models. Without flexible and customizable constraints, counterintuitive or unexplainable results may appear, and some groups of customers may be disadvantaged when determining their credit risk.

In this paper, we use our toolbox that features an ML algorithm leveraging modified logistic regression with predefined constraints imposed via automated supervised binning and variable selection. We use this algorithm to build a benchmark scorecard model that is interpretable by design. We then compare this model with three challenger models built using other classifiers, decision tree, random forest and gradient boosting methods, in terms of model performance and interpretability. To assess the challenger models' interpretability, we use the toolbox to identify customer characteristics that do not yield desired patterns and hence violate constraints.

Our results demonstrate that there is no significant difference in performance between interpretable benchmark model and challenger classifiers models. Using different size datasets for personal loans and credit cards, we find that the benchmark model performance is overall slightly inferior to challenger models in model fit and discriminatory power. We show that while the gradient boosting and random forest models can provide superior fit to the benchmark model, they do it at the cost of violating many constraints.

The remainder of this paper is organized as follows. In section two, we survey the main applications of machine learning methods in scorecard-building. We then describe the methodology behind our algorithmic toolbox, which includes binning and variable selection for the benchmark model, and constraint violation assessment for the challenger models. In section three, we assess the performance of benchmark and challenger models using our toolbox in terms of their performance and interpretability.

Challenges of scorecard-building

The scorecard models are designed to rank-order customers by condensing the variety of variables into a single score. Scorecard types differ by the target variable, such as application, behavioural and collection scorecard. The models typically use the data for customer and product characteristics, while alternative information such as transactional data, telecommunication, rental and utilities variables can add more insight and predictive value. Such datasets vary by size and structure.

Industry embraces ever-expanding and improving ML techniques at various stages of scorecard development, from data preparation and variable selection to model building, optimization and monitoring. Since computation power has increased dramatically over the last decade, advanced machine learning methods such as gradient boosting, neural networks and random forests have made their way into credit-scoring models. These methods have demonstrated their superiority in the speed of the scorecard-building process and often predictive power compared to the traditional logistic regression approach, see Efron & Hastie (2016) and Alpaydin (2020), among others. More complex classifier methods have proven to be especially superior in large unstructured datasets with many predictors. Chart 1 summarises applications of machine learning in scorecard-building.

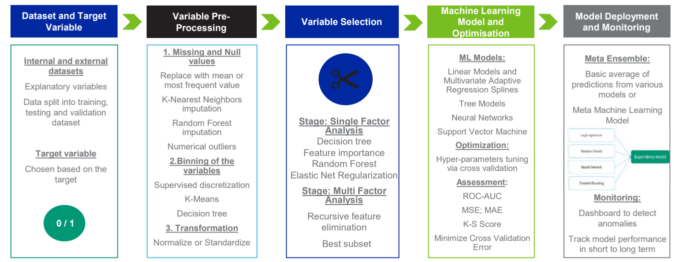
However, many advanced machine learning methods suffer from a lack of tractability and interpretability of model structure and predictions. The "interpretable" trend in ML model development has been gaining more attention, see, for example, Gilpin et al (2018) and Rudin (2019). Although there are some methods that enable us to peek into the black box, there is still no consensus on how to assess the interpretation quality.[1] As the interpretability translates into a set of constraints, algorithmic solutions for more complex ML methodologies are challenging. These issues have been recognized by industry and regulators worldwide, who called for the responsible use of ML to ensure the principles of fairness, ethics, accountability and transparency when assessing customers' credit risk.

Recognizing the need for a scorecard to be interpretable, transparent, and able to withstand regulatory scrutiny, we have designed an automated algorithmic toolbox. The scorecard toolbox includes tools for data analysis, model development and assessment, model validation, model refinement, scoring, model-monitoring, and strategy-setting. At the core of this toolbox are the binning and the variable selection algorithms that solve optimization problems subject to user-specified constraints required in credit applications.

The scorecard toolbox allows us to use traditional and enhanced regression methodologies as well as alternative ML methods. Such an approach allows us to focus on key decisions, while relegating the tedious, complex tasks of binning, variable selection, and the constraints violations assessment to automated algorithms.

Chart 1: ML in Scorecard-Building

5 Stages



Source: Moody's Analytics

Binning algorithm

Binning is a first step in scorecard model development. It transforms the values of various types of potential predictors into several groups, known as bins, according to specified criteria. Binning is applied to numerical data such as customer age or income, categorical data such as loan purpose or property type, and ordinal data that has defined ordering, such as customer education or employment status. The result of binning is a set of "binned" variables for the next step in model development, the variable selection.

The key advantages of binning include

- **Simplicity and business tractability.** Binning is used to simplify the model predictors by creating groups that have expected patterns and relationships with the target variable. For instance, low-income customers are expected to have higher default rates than higher-income customers. Hence, it makes sense to split the numerical income values into several bins. Including binned variables allows us to evaluate only a few logical conditions to calculate the score, instead of calculating the score for each possible combination of predictor values.

- **Flexibility to incorporate constraints.** Binning can be formulated with various types of constraints. These constraints include binning size constraints, logical patterns, business expectations, and compliance with equal credit opportunity legal acts for customer age or gender. For example, a monotone or quadratic relationship between binned predictors and default rate can be incorporated as a constraint when splitting variable values into bins.

- **Capture non-linear relationships.** Binning allows us to capture non-linearities in a data-driven way, without making restrictive parametric assumptions. For example, account age may have a non-linear relationship with the default rate.

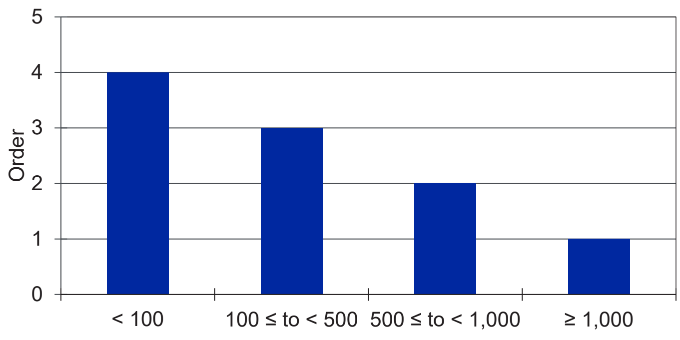
- **Model accuracy by handling outliers and missing values.** Binning mitigates the impact of outliers and missing values by grouping observations. Grouping of similar attributes with similar predictive strengths increases the model's accuracy. For example, the procedure extracts information from such observations into a separate bin and uses it to predict the target variable.

In practice, binning procedures vary depending on data and model characteristics. Binning can be based on expert opinion, utilize unsupervised or supervised algorithms with quantitative optimization techniques, or use a combination of these. Typically, many manual interventions and visual assessments of the binning solution's quality are required.

1 For example, Lundberg and Lee (2017) developed Shapley Additive Explanations to interpret the output of machine learning models, while Carvalho et al (2019) provide a review of machine learning models' interpretability.

Chart 2: Constraints for Ordinal Data

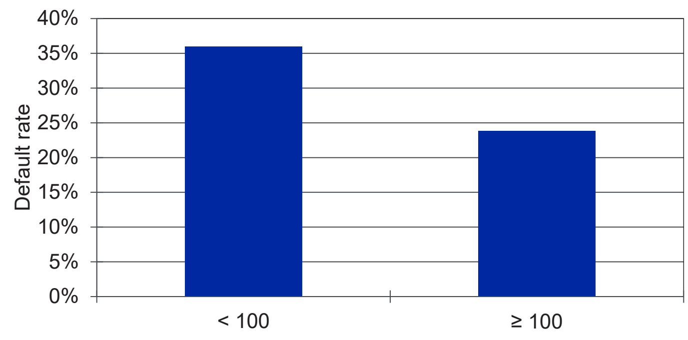
Assumption



Source: Moody's Analytics

Chart 3: Constraints for Ordinal Data

Solution



Source: Moody's Analytics

The main idea of supervised binning is to find optimal cut-off points to define bins subject to various types of constraints. Some constraints are required to ensure each bin strikes a balance between being "wide enough" and "narrow enough" by having distinctly different risk characteristics with minimum information loss. Examples include controlling the number of bins, the number of observations in each bin, and non-overlapping confidence intervals for default rates of each bin. In addition, a critical aspect of binning is the enforcement of various types of constraints representing requirements that certain patterns must emerge when calculating the scores.

Our toolbox automates the tedious aspects of supervised binning, allowing the analyst to specify options and preferences. The supervised binning algorithm solves an optimization problem with user-defined constraints, while controlling the number and discriminatory power of resulting bins. The procedure significantly reduces the time costs of generating predictive characteristics.

The key ingredients of our supervised binning algorithm include

- **Maximize binned variable's predictive power.** User-defined performance metrics such as information value, Gini and chi-square statistics to assess variable's predictive power.

- **Comply with constraints.** User-defined constraints and label-ordering for ordinal variables are incorporated into the optimization algorithm. These constraints represent expected trends in the default rates across bins. The toolbox implements both monotone (such as decreasing, increasing) and non-monotone (such as u-shaped, hump-shaped) types of relationship between the binned variable and target variable. Moreover, the toolbox allows for the incorporation of the constraints for ordinal variables based on the user-provided order of labels.

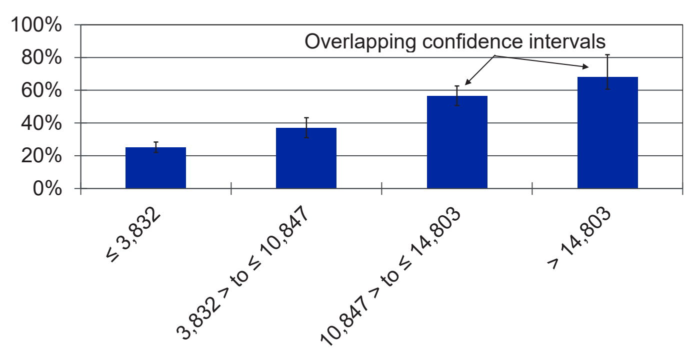
  Charts 2 and 3 illustrate an example of implementing constraints for ordinal data for available savings. Chart 2 shows the preferred order of the categories specified by the user: The default rates follow a non-increasing trend across categories as lower savings are associated with worse credit quality. Chart 3 demonstrates the solution by the supervised binning algorithm obtained in line with this imposed assumption. The categories "100 ≤ to < 500", "500 ≤ to < 1,000" and "≥ 1,000" are merged, but the desired ordering is preserved as the trend in default rates is indeed non-increasing.

- **Control number, size and discriminatory power for selected bins.** Users can control the number and size of bins as well as specify the threshold level of confidence interval to assess the discriminatory power of the selected bins. To improve the quality of binning, the algorithm considers not only the point estimate but also a confidence interval for the default rate of each bin. For example, when confidence intervals for each bin overlap, the chosen bins may not have enough power to discriminate between defaulted and non-defaulted observations.

  As illustrated in Charts 4 and 5, the algorithm combines such bins to achieve an optimal solution. Chart 4 shows an example binning solution with overlapping confidence intervals for the 3rd and 4th bins. Chart 5 demonstrates the toolbox solution that satisfies the constraint on non-overlapping intervals. In this example, this is achieved by merging intervals "10,847 to 14,803" and ">14,803" into ">10,847".

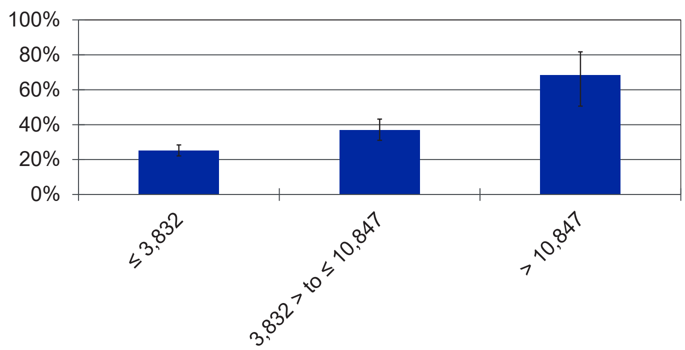
Chart 4: Confidence Intervals Merging

Default rate



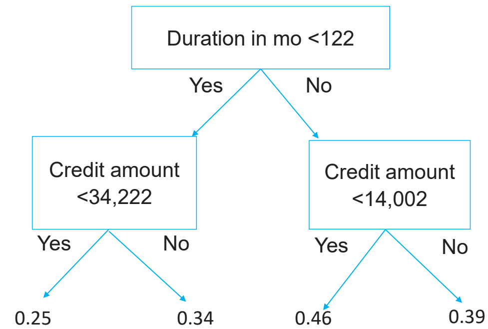
Source: Moody's Analytics

Chart 5: Confidence Intervals Merging



Source: Moody's Analytics

Chart 6: A Simple Decision Tree Example



Source: Moody's Analytics

Variable selection algorithm

As a next step of the model-building process, the variable selection algorithm is used to identify the characteristics to include into the regression model. The binned candidate variables obtained at the previous binning stage can be used as inputs into the variable selection procedure to define the model specification.

In practice, various methodologies are used for the variable selection in various types of risk models. Brute-force algorithms exhaustively evaluate all possible combinations of candidate predictors to find the best subset. These algorithms can be modified to incorporate constraints. Dynamic credit risk models with linkages to macroeconomic as well as portfolio characteristics are good examples when such procedures work well, see Licari, Loiseau-Aslanidi and Vikhrov (2017).

Alternative variable selection procedures are preferred when the number of candidate characteristics and number of observations is so large that it makes the exhaustive search's computational cost prohibitively high. Stepwise algorithms such as forward stepwise have the advantage of relatively high execution speed, as they rely on a sequence of nested models as opposed to the brute-force exhaustive search. Not surprisingly, forward stepwise regression is a workhorse variable selection procedure in credit risk scoring.

Nevertheless, classical stepwise methods require enhancements to improve their efficiency and applicability for scorecards. First, the stepwise algorithm is not robust to variable ordering. The order in which variables enter the model has a significant impact on the final model that may result in overfitting, dependence on training sample selection, or that may eliminate variables that would provide additional information and improve model performance, see Altman & Anderson (1989) and Audrino & Kanus (2016) among others. Second, stepwise regression does not consider the possible correlations between the variables. Nor does this algorithm consider constraints such as logical patterns based on business experience, industry trends, or legally required relationships.

In our toolbox, we enhance the stepwise algorithm by offering capabilities to specify user preferences on dependencies. Such constraints may include the expected relationships between characteristics and the target variable, statistical significance of the variables, and maximum allowed value of pairwise correlation. These constraints may be imposed either on coefficients' estimates signs or their order. After the model is built, the validation is performed, and an iterative model refinement algorithm sequentially excludes variables that do not comply with the constraints from a list of initial potential drivers.

Assessing constraints violation for challenger models

The scorecard model designed using the steps outlined above is designed to satisfy user-defined constraints imposed at the supervised binning and the variable selection stages. In contrast, decision tree, random forest and gradient boosting challenger models need an additional analysis to assess models' tractability. Our toolbox provides several options to identify the constraints violations for these machine learning models.

In the case of decision tree, it is feasible to extract all tree nodes. We calculate the average probability of default for each variable interval determined by tree cut points. If the variable appears at different tree branches several times, the average default probability is calculated based on all internal and leaf nodes, taking into consideration interaction terms. This procedure is analogous to evaluating the type of the relationship in the case of binning and is, therefore, straightforward to use for evaluation of the constraint's violation. A simple tree example illustrating the procedure is shown in Chart 6.

In the case of random forest and gradient boosting, the extraction of all tree nodes and splits is not the best solution because of their complicated model structure. We rely on the Shapley values approach used to assess the marginal contributions of various drivers into predicted probability of default values. Because of the potentially very large number of cut points for continuous variables, the evaluation of trend monotonicity based on the average default rates for each bin may not be applicable.

To facilitate comparability of the results with a logistic regression model, we focus on evaluation of the constraints only for ordinal variables and calculate average Shapley values for each category. Additionally, the toolbox provides a standard heat map of Shapley values for various realizations of each driver.

Models Assessment

Data and methodology

To assess and compare the performance of several ML models, we conduct an empirical study using two datasets that differ by product type, geography and size. Both datasets cover consumer credit portfolios from the UCI Machine Learning Repository[2], which is frequently used in studies on performance evaluation of machine learning and data mining algorithms. The first dataset covers a German fixed-term portfolio for personal loans, while the second dataset covers a credit cards portfolio in Taiwan (see Table 1).

We begin by splitting each dataset into development (train) and holdout (test) in a standard proportion 70:30. To mitigate a sample-dependency bias for model performance measures, especially for the German dataset consisting of only 1,000 observations, we generate 100 train and test subsets realizations without replacement.

For each realization of the train dataset, we build four alternative models. The first model is a benchmark built using the algorithmic supervised binning and modified weight-of-evidence logistic regression with example constraints presented in Table 2. Next, we use decision tree, random forest and gradient boosting of decision trees with least-squares loss function as three challenger models.

For the latter models, it is crucial to properly tune hyper-parameters to prevent overfitting. We tune hyper-parameters through stratified k-fold cross validation with application of exhaustive grid search over various parameter combinations to maximize the average accuracy ratio on validation subsamples. For the RF model, the procedure optimizes the maximum depth of trees and number of estimators. For the GB model, the set of optimized parameters is broader, and along with maximum depth and number of trees it includes the learning rate and subsample size to be selected for the estimation of each tree.

2 This is a real-life credit scoring dataset publicly available at the UCI repository at http://kdd.ics.uci.edu/.

Table 1: Summary of Dataset

Country	Germany	Taiwan
Product type	Personal loans	Credit cards
Number of observations	1,000	30,000
Number of characteristics	24	23
Number of defaults	300	6,636

Sources: UCI Machine Learning Repository, Moody's Analytics

Model performance and interpretability assessment

We use standard measures to evaluate the models' performance, and we assess models' interpretability by looking into constraints violations. The accuracy of the model fit is measured by the Brier score, while the discriminatory power is assessed through Gini or the area under the curve (see Table 3).

We observe that enforcing constraints on the benchmark model has little impact on the performance, and for brevity we do not report the results of the benchmark model without constraints. Moreover, the benchmark regression model with supervised binning and constraints demonstrates similar

Table 2: Selected Example Constraints for Model Interpretation Evaluation

Variable	Product type	Variable type	Trend	Order of labels for ordinal variables
Duration in mo	Personal loans	Numerical	Monotonic	
Savings account amount	Personal loans	Ordinal	Negative	< 100, 100 <= ... < 500, 500 <= ... < 1000, >= 1000
Credit amount	Personal loans	Numerical	Positive	
Present employment since (employment longevity)	Personal loans	Ordinal	Negative	Unemployed, ... < 1 yr, 1 <= ... < 4, 4 <= ... < 7, >= 7 yrs
Installment rate	Personal loans	Numerical	Monotonic	
Other debtors	Personal loans	Ordinal	Negative	None, co-applicant, guarantor
Present residence since	Personal loans	Numerical	Negative	
Housing type	Personal loans	Ordinal	Negative	For free, rent, own
Number of existing credits	Personal loans	Numerical	Positive	
Job	Personal loans	Ordinal	Negative	Unemployed, unskilled, unskilled resident, skilled employee, official, management, self-employed, highly qualified employee, officer
Number of people being liable	Personal loans	Numerical	Negative	
Telephone	Personal loans	Ordinal	Positive	Yes, registered under the customer's name, none
Amount of given credit	Credit cards	Numerical	Positive	
Education	Credit cards	Ordinal	Negative	NA, high school, others, university, graduate school
Past payment status in Apr-Sep 2005	Credit cards	Ordinal	Positive	
Amount of bill statement (balance) in Apr-Sep 2005	Credit cards	Numerical	Monotonic	
Amount of previous payment in Apr-Sep 2005	Credit cards	Numerical	Monotonic	

Source: Moody's Analytics

Table 3: Summary of Model Performance

	Personal loans		Credit cards	
	Gini	Brier score	Gini	Brier score
WOE logistic with supervised binning with constraints	0.519	0.175	0.525	0.139
Decision tree	0.417	0.198	0.41	0.14
Random forest	0.513	0.177	0.567	0.132
Gradient boosting	0.544	0.173	0.562	0.133

Source: Moody's Analytics

performance to the challenger ML models.[3] The GB model performs slightly better for personal loans, while for credit cards the RF model outperforms the others. The DT model's performance is inferior for both product types.

To evaluate the models' interpretability, the violation of constraints is evaluated for the challenger DT, RF and GB models (see Table 4). The benchmark model satisfies all the constraints by design. In the case of personal loans, the constraints are violated for six out of 10 variables that appear in the DT model. For example, the credit amount and the number of existing credits do not have the expected positive relationship with the default rate. Similarly, the employment longevity and job description ordering are counterintuitive, resulting in the unemployed customers having lower default rates than those with years of employment. For the RF and GB models, all the constraints are violated, with the employment-related variables having the highest percentage of violation cases.

In the case of credit cards, the selected variables represent the most recent (September and August) past payment status and previous payments. The rest of the lags are not selected by the models because of the absorbing properties of the most recent observations. As expected, the constraints are not violated in the DT model because of the relatively shallow trees produced by the algorithm, while the constraints are violated for the RF and GB models.

The illustration of constraints violations for employment status is depicted in Charts 7-10. In contrast to the benchmark model with supervised binning, where lower default rates are associated with longer duration of employment, the models using the DT, RF and GB methods show counterintuitive relationships. The DT model predicts a lower default rate for the categories with unemployed and shorter employment duration than for categories with longer employment duration. Both the RF and GB models predict lower default rate for the "Unemployed" versus the 1-year employment duration categories. Additionally, the RF model predicts a somewhat lower default rate for four to seven years than for seven years or more.

3 We compare performance based on the accuracy of their predictions on the test dataset. Model performance on train datasets cannot be used for an objective model evaluation since that dataset was used for model development.

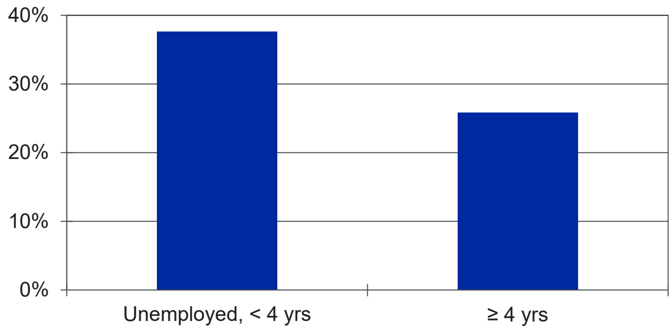
Table 4: Frequency of the Example Constraints Violation

Constraint variables	Data	Variable type	% of appearance, DT	% of violated cases, DT	% of violated cases, RF	% of violated cases, GB
Duration in mo	Personal loans	Numerical	100	0		
Credit amount	Personal loans	Numerical	95	26.32	-	-
Present employment since	Personal loans	Ordinal	50	16	92	93
Other debtors	Personal loans	Ordinal	32	0	62	52
Savings account amount	Personal loans	Ordinal	26	0	66	39
Installment rate	Personal loans	Numerical	22	0	-	-
Present residence since	Personal loans	Numerical	25	92	86	97
Housing type	Personal loans	Ordinal	0	-	33	52
Job	Personal loans	Ordinal	5	60	91	91
Telephone	Personal loans	Ordinal	7	42.86	64	31
Number of existing credit	Personal loans	Numerical	3	100	-	-
Number of people being liable	Personal loans	Numerical	0	-	-	-
Education	Credit cards	Ordinal	0	-	0	100
Past payment status in Sep 2005	Credit cards	Ordinal	100	0	100	100
Past payment status in Aug 2005	Credit cards	Ordinal	100	0	100	100
Past payment status in Jul 2005	Credit cards	Ordinal	0	-	100	100
Past payment status in Jun 2005	Credit cards	Ordinal	0	-	100	100
Past payment status in May 2005	Credit cards	Ordinal	0	-	100	100
Past payment status in Apr 2005	Credit cards	Ordinal	0	-	100	100
Amount of previous payment in Sep 2005	Credit cards	Numerical	6	0	-	-
Amount of previous payment in Aug 2005	Credit cards	Numerical	28	0	-	-
Amount of previous payment in Jul 2005	Credit cards	Numerical	1	0	-	-
Amount of previous payment in Apr-Jun 2005	Credit cards	Numerical	0	0	-	-
Amount of bill statement (balance) in Apr-Sep 2005	Credit cards	Numerical	0	0	-	-

Source: Moody's Analytics

Chart 7: Default Rates vs. Employment

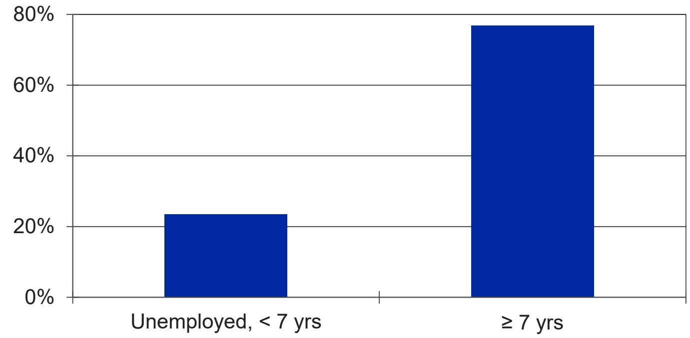
Default rate, constraint logit



Source: Moody's Analytics

Chart 8: Default Rates vs. Employment

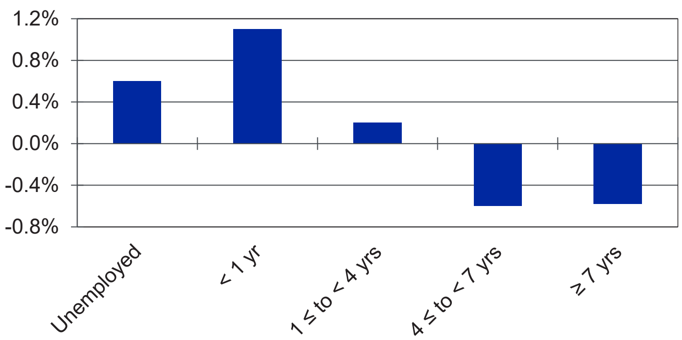
Default rate, decision tree



Source: Moody's Analytics

Chart 9: Shapley Values vs. Employment

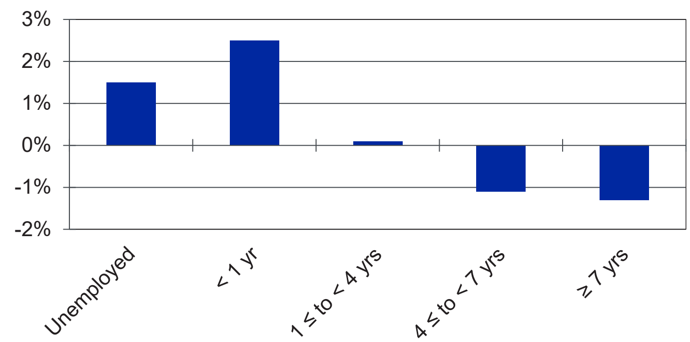
Avg Shapley value, random forest



Source: Moody's Analytics

Chart 10: Shapley Values vs. Employment

Avg Shapley value, gradient boosting



Source: Moody's Analytics

Conclusion

Using our automated toolbox, we designed and compared several models in terms of their performance and interpretability. The considered models include the benchmark model leveraging modified logistic regression with supervised binning, and three challenger models using the decision tree, random forest and gradient boosting methods.

Models' interpretability is represented by a set of constraints. The key feature of the used toolbox is the broad type of user-defined customizable constraints used in supervised binning and variable selection algorithms for the benchmark model. The toolbox is also used to assess the challenger ML models' results by looking into constraints violations.

We demonstrated that our benchmark model achieves somewhat similar performance to the challenger ML models, while being easily interpretable and satisfying imposed constraints by default. Although the gradient boosting and random forest models slightly outperform the benchmark model, this is achieved at the expense of the high risk of constraints violation. The challenger models produce counterintuitive results for some model characteristics, making further model refinements necessary before the model is implemented for decision-making.

Our algorithmic tools to build the benchmark model allow for a reasonable compromise between the automation to reduce manual intervention and minimize the time and costs of model development, and the model interpretation that is critical in credit scoring applications.

References

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Altman, D., & Anderson, P. (1989). Bootstrap investigation of a Cox regression model. Statistics in Medicine, 8(7), 771-83.

Anderson, R. (2007). The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford University Press.

Audrino, F., & Kanus, S. (2016). Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics. *Econometric Reviews*, 35, 1485-1521.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv preprint arXiv:1806.00069.

Licari, J. M., Loiseau-Aslanidi, O., & Vikhrov, D. (2017). Dynamic Model-Building: A Proposed Variable Selection Algorithm. Moody's Analytics Risk Perspectives. Volume IX.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Siddiqi, N. (2017). Intelligent credit scoring: Building and implementing better credit risk scorecards. John Wiley & Sons.

About the Authors

Olga Loiseau-Aslanidi is the head of risk modelling of Economics and Business Analytics APAC, based in the Singapore office. She manages a team of economists and risk modelers in Prague, Shanghai and Sydney who design models for forecasting and simulation, with an emphasis on stress-testing for three key areas: macroeconomic models, market risk, and credit portfolio risk. During her time at Moody's based in Europe and now in Asia, Olga has led consulting projects with major banks and other financial institutions worldwide focusing on stress-testing, including CCAR, EBA, PRA as well as IFRS 9, IRRBB and IRB model design and implementation. She is directly involved in the research and implementation of Moody's Analytics risk management solutions for market risk and retail credit risk modelling, and often speaks at credit events and economic conferences worldwide, communicating the team's research and methodologies to the market. Before joining Moody's Analytics, she worked for the Academy of Sciences of the Czech Republic and at a consultancy firm focused on macroeconomic forecasting and analysis of emerging market economies. She has published several academic articles and has been teaching graduate courses in economics, statistics and econometrics. She holds a PhD and an MA in economics and econometrics from Charles University (CERGE-EI), following studies for an MSc in mathematics and a BSc with honors in applied mathematics.

Natchie Subramaniam Thiagarajah is an associate director at Moody's Analytics and a member of the Credit Analytics group. Subra is involved in developing and validating credit risk models. Before joining Moody's Analytics, Subra was a senior associate at Discover Financial Services, where he did model validation and research. Subra holds a PhD in economics and an MS in agricultural economics from the University of Arizona. Subra completed his BS in agricultural economics at the University of Peradeniya in Sri Lanka.

Vera Tolstova is an economist in the Prague office. She is responsible for leading advisory projects involving credit and market risk and IFRS 9 and stress-testing model development and validation with banks and international financial institutions. Additionally, Vera and her team work on developing methodologies and numerical algorithms for credit risk PD, LGD, EAD and IRB modelling involving automatized binning, variable selection and machine learning methods for credit risk scorecards, hybrid scorecards combining both quantitative and expert judgment scores as well as forecast performance analysis according to IFRS 9 methodology. Prior to joining Moody's, Vera was working as a junior researcher at the Center of Economic Research and Graduate Education (CERGE-EI) in Prague and as an analyst at one of the leading Russian pharmaceutical distributors. She received her MA in economics from CERGE-EI in 2014, following master's and bachelor's degrees in economics from Novosibirsk State University, Russia, and is currently working on her PhD dissertation thesis with a focus on dynamic macroeconomics with heterogeneous agents, public policy and family economics.

About Moody's Analytics

Moody's Analytics provides financial intelligence and analytical tools supporting our clients' growth, efficiency and risk management objectives. The combination of our unparalleled expertise in risk, expansive information resources, and innovative application of technology helps today's business leaders confidently navigate an evolving marketplace. We are recognized for our industry-leading solutions, comprising research, data, software and professional services, assembled to deliver a seamless customer experience. Thousands of organizations worldwide have made us their trusted partner because of our uncompromising commitment to quality, client service, and integrity.

Concise and timely economic research by Moody's Analytics supports firms and policymakers in strategic planning, product and sales forecasting, credit risk and sensitivity management, and investment research. Our economic research publications provide in-depth analysis of the global economy, including the U.S. and all of its state and metropolitan areas, all European countries and their subnational areas, Asia, and the Americas. We track and forecast economic growth and cover specialized topics such as labor markets, housing, consumer spending and credit, output and income, mortgage activity, demographics, central bank behavior, and prices. We also provide real-time monitoring of macroeconomic indicators and analysis on timely topics such as monetary policy and sovereign risk. Our clients include multinational corporations, governments at all levels, central banks, financial regulators, retailers, mutual funds, financial institutions, utilities, residential and commercial real estate firms, insurance companies, and professional investors.

Moody's Analytics added the economic forecasting firm Economy.com to its portfolio in 2005. This unit is based in West Chester PA, a suburb of Philadelphia, with offices in London, Prague and Sydney. More information is available at www.economy.com.

Moody's Analytics is a subsidiary of Moody's Corporation (NYSE: MCO). Further information is available at www.moodysanalytics.com.

About Moody's Corporation

Moody's Analytics is a subsidiary of Moody's Corporation (NYSE: MCO). MCO reported revenue of $4.8 billion in 2019, employs more than 11,000 people worldwide and maintains a presence in more than 40 countries. Further information about Moody's Analytics is available at www.moodysanalytics.com.