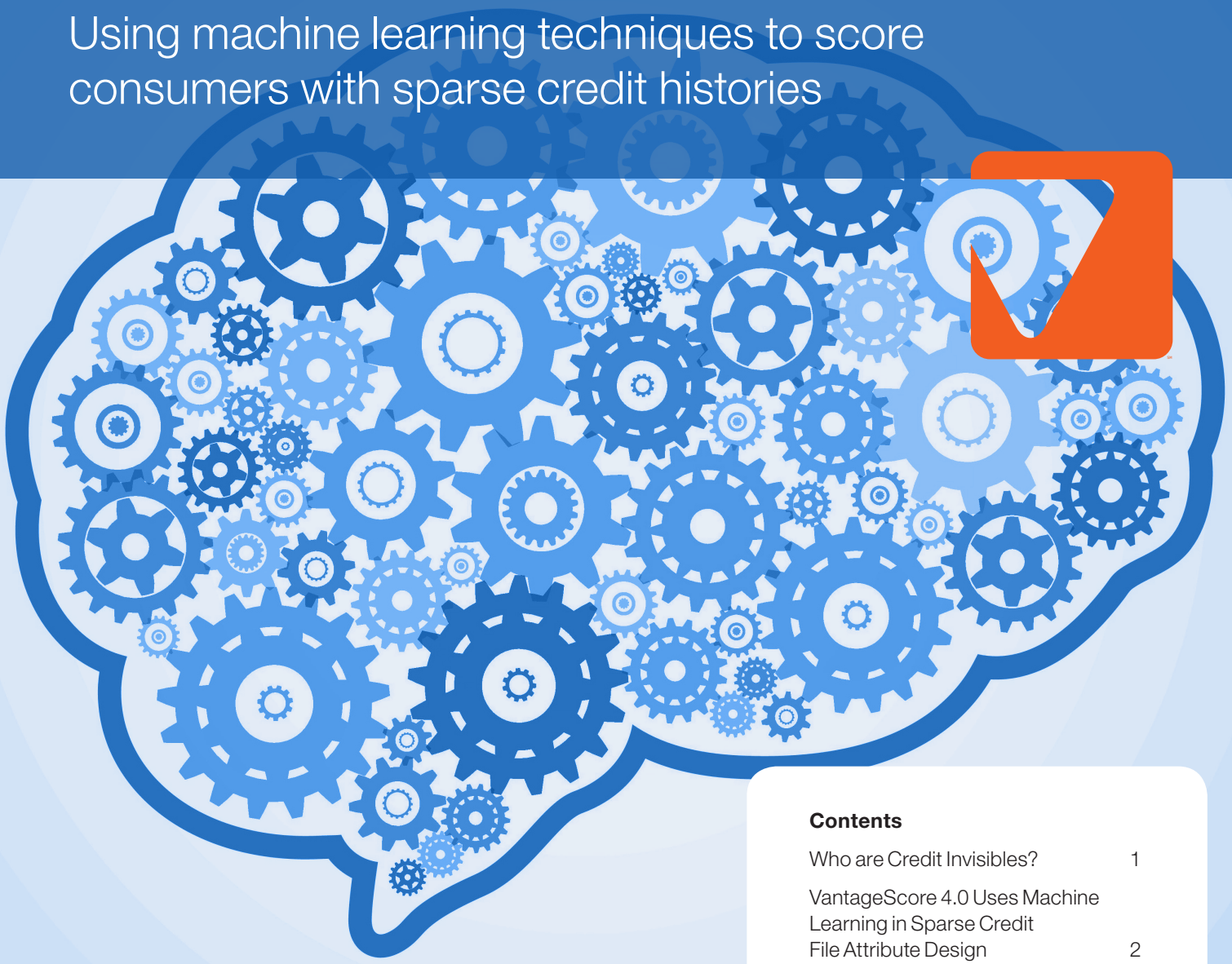
OCTOBER 2017

Scoring Credit Invisibles

Using machine learning techniques to score consumers with sparse credit histories



Contents

VantageScore®
a higher level of confidence

Scoring Credit Invisibles

Using machine learning techniques to score consumers with sparse credit histories

Over the past months, much has been made about the potential for using machine learning techniques to improve the analysis of risks of consumer lending. Much of the discussion tends to be hypothetical or concerns applications that are outside the realm of credit decisioning. The development of VantageScore 4.0 showcases how these technologies can be harnessed in a way that marries both the latest innovations and current compliance considerations. By using a score that incorporates these techniques, lenders in turn can take advantage of the most recent model improvements with relative ease.

Indeed, VantageScore 4.0 harnesses improved performance by incorporating a machine learning attribute design approach known as random forest methodology. This research study will review the population segments that benefit from this innovative modeling method, as well as how and why this approach facilitates increased predictive ability in VantageScore 4.0 for the so-called credit invisibles, i.e., those consumers that conventional credit scoring models ignore.

SUMMARY

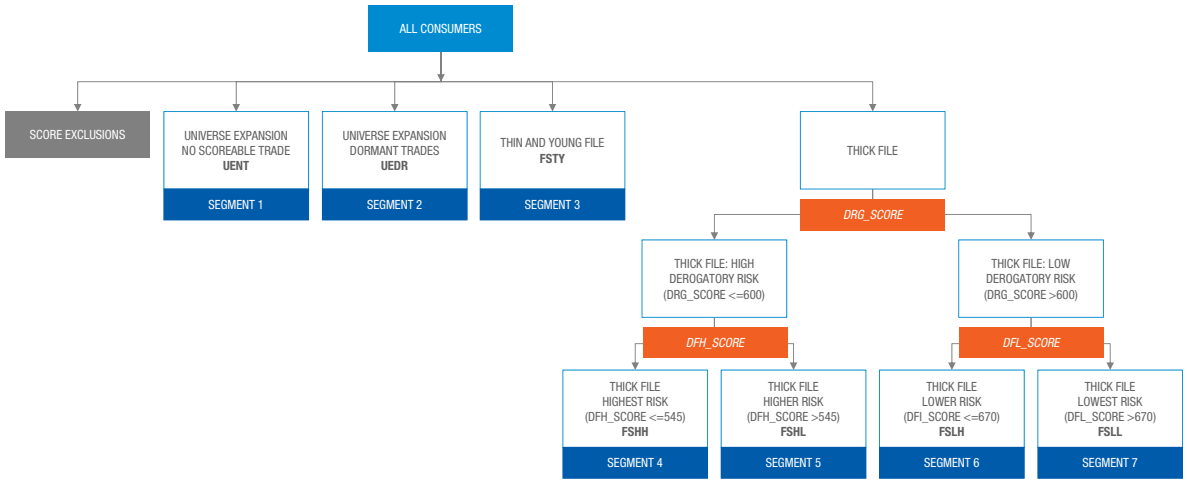
- In VantageScore 4.0, machine learning drives enhancements to scorecard development for sparse credit consumers, often called credit invisibles because conventional models are unable to score these consumers.
- Version 4.0 uses a random forest classifier approach to generate performance improvements. This involves a three-step process that generates nearly 50,000 predictive behavioral nodes, aggregates these into a superset of high performing nodes, and lastly converts these nodes, using regulatory and business rules, into the conventional structured attributes that are typically used in scoring models.
- This approach resulted in more than a 10 percent lift in performance of the Dormant segment and more than a 30 percent lift to the No Trade segment as compared to the performance generated by conventional model design methods.

WHO ARE CREDIT INVISIBLES?

Credit Invisibles are consumers with atypical, often sparse credit files. These consumers may not have a trade that is at least six months old or they may not have had an update to their credit file in the last six months. Consequently, these consumers cannot be scored by conventional credit scoring models that need sufficient volume of profile information in order for those conventional models to generate a score.

In addition to scoring mainstream consumers (i.e., those with typical credit file profiles), VantageScore models are intentionally designed to score this credit invisibles population. In the development of VantageScore 4.0, two specific scorecards were formulated to score this population. The Dormant segment scores consumers who have scoreable trades but have had no update to their credit file in the last six months (Figure 1, Segment 2). The No Trade segment scores consumers with no scoreable trades on their file, but who have collections and public records (Figure 1, Segment 1). Consumers with no trades older than six months are scored in the Thin and Young segment (Figure 1, Segment 3). Consumers with “only inquiries” are not scored by VantageScore 4.0.

Figure 1: VantageScore 4.0 Segmentation



VANTAGESCORE 4.0 USES MACHINE LEARNING IN SPARSE CREDIT FILE ATTRIBUTE DESIGN

When seeking to improve the performance of credit score models, two approaches are typically used: to incorporate additional data (e.g., rent, utility and cell phone information) into the model or to use enhanced mathematical techniques to describe the predictive relationships within the existing behavioral credit data. By using machine learning, VantageScore is able to leverage both approaches, to take advantage of additional data that may be reported in the consumer's primary credit file as well as to implement innovative modeling techniques.

In VantageScore 4.0, machine learning was used to augment the development of credit data attributes for a subset of the population (i.e., those with sparse credit histories). In the past, this subset was particularly difficult to assess when using uni- or bi-dimensional attributes. For example, when assessing consumers with dense credit files, uni-dimensional attributes, such as the number of inquiries or the amount of a mortgage balance, are predictive of lower risk while a higher number of inquiries or a higher mortgage balance empirically indicate higher risk.

However, under sparse data conditions, such 'simple' attributes are often not sufficiently sensitive to predict accurately consumer risk of default. By combining multiple behavioral dimensions into a specific configuration, or node, of behaviors, a model is able to identify substantial and additional risk assessment for such consumers. These nodes will then be converted into traditional, structured attributes for consideration within the standard stepwise discriminant analysis process to determine the optimal attributes for the scorecard.

HOW MACHINE LEARNING IMPROVES THE PREDICTIVENESS OF VANTAGESCORE 4.0

Normally, model attributes incorporate only one or two dimensions from the behavioral credit data that is available. The random forest approach allows multiple behavioral dimensions to be randomly combined into highly predictive nodes that may then be structured into intuitive and logical attributes. Consider the following example (in Figure 2) of a final attribute, which indicates that consumers with higher balances on newer collection accounts and who are actively seeking credit present a greater default risk than those consumers who have older accounts, lower balances and are not actively seeking credit. Note: the desired monotonicity in the default rate (90+ days past due) that aligns with simple attribute behaviors. For example, within a particular balance tier, as the accounts become older, the risk decreases, indicating that consumers have had more time to resolve the debt. As is typical, increasing inquiries indicates higher risk. Similarly, within the same time period, higher balances indicate higher risk. This configuration of insights only emerges after drilling down on the appropriate balance, age and inquiry limits. If only one or two of the attributes had been evaluated, this risk profile would not have emerged. Figure 2: Multi-dimensional structured attribute example.

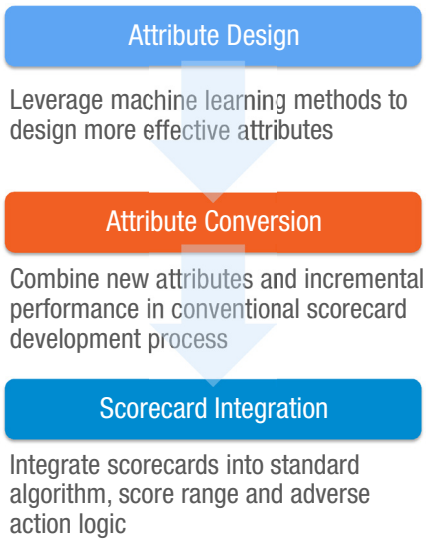
Figure 2: Medical Collection Accounts Default Rate Profile by Balance, Age with Inquiry Volume [Unpaid, >6 months age]

Balance $	Age	Number of inquiries 1	2	3+
<=1,000	7-12 mos	32.6%		
	12-36 mos	30.4%		
	36+ mos	27.7%	28.7%	32.1%
>1,000	7-12 mos	32.7%	40.5%	
	12-36 mos	32.1%	34.6%	
	36+ mos	30.3%	34.5%	

DEVELOPMENT PROCESS

The three-step Development Process (Figure 3) initially designs and generates 50,000 random behavioral nodes that are then evaluated and aggregated into a superset of high performing nodes in order to provide a benchmark for the optimal performance for the scorecard. Next, these high-performing nodes are converted into conventional structured modeling attributes. Finally, these attributes are integrated into the traditional scorecards used within generic risk scoring models. The process allows for statistical power and analytical rigor while enabling regulatory compliance.

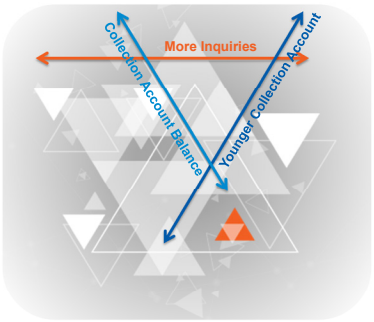
Figure 3: Scorecard Development Process



Step 1: Attribute Design

Attribute design begins with the development of random forest trees that are comprised of unstructured behavioral nodes. Up to 500 trees are generated per scorecard. Each tree is considered to be thousands of nodes, made up from random combinations of as many as 100 behavioral credit dimensions, such as revolving credit balance, available credit amount and number of months reported. Each dimension is designed to allow the full range of permissible values for the behavior. Subsets of these ranges of values across multiple dimensions are randomly selected and combined to construct the node (Figure 4). The performance (pay and default rates) for the node is calculated. Multiple nodes are generated within each tree in order to fully capture the pay and default performance.

Figure 4: Node example



Node_1_12 DESCRIPTION:

IF COLLEXT_BAL is greater than or equal to 0 and is less than 88

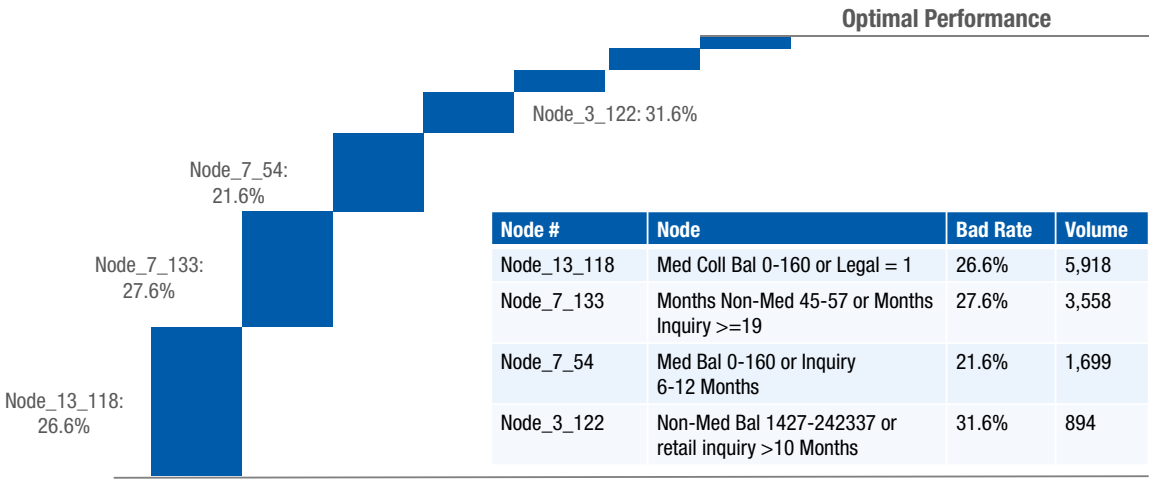
AND MONTHS_REPORTED is greater than or equal to 5 and is less than or equal to 11

AND AVAILABLE_CREDIT_AMT is greater than or equal to 0 and is less than 500,

THEN there is a 74% chance that consumer will pay and a 26% chance that consumer will default.

As many as 50,000 nodes are generated. The highest performing-highest volume nodes are identified and combined to provide an estimate of the optimal predictive performance (Figure 5). The goal now is to convert these unstructured high-performing nodes into structured attributes that capture as much of the optimal performance as possible.

Figure 5: Node-driven optimal scorecard performance



Node #	Node	Bad Rate	Volume
Node_13_118	Med Coll Bal 0-160 or Legal = 1	26.6%	5,918
Node_7_133	Months Non-Med 45-57 or Months Inquiry >=19	27.6%	3,558
Node_7_54	Med Bal 0-160 or Inquiry 6-12 Months	21.6%	1,699
Node_3_122	Non-Med Bal 1427-242337 or retail inquiry >10 Months	31.6%	894

Step 2: Attribute Conversion

A supervised process of combining nodes occurs to merge relevant nodes and key behaviors into a structured attribute (Figure 6). This process is more of an art, using innovation to incorporate standard regulatory conditions into the new model attributes. The priority within this process is to prevent performance loss, while combining nodes into logical and intuitive attributes that meet adverse action requirements.

Figure 6: Example - node aggregation to develop a structure attribute

Sample Nodes

Name	Node Description
Node_1_85	If COLLEXT_BAL is greater than or equal to 200 and is less than 788 and MONTHS_REPORTED is greater than or equal to 6 and is less than or equal to 11 and AVAILABLE_CREDIT_AMT is greater than or equal to 0 and is less than 500 then there is a 73.6567 percent chance that aaa_mod_var will be 0 and a 26.3433 percent chance than aaa_mod_var will be 1.
Node_1_12	If MAX_RATE is equal to 0 and COLLEXT_BAL is greater than or equal to 0 and is less than 88 and NUM_INQ is greater than or equal to 2 and is less than or equal to 6 then there is a 73.6567 percent chance that aaa_mod_var will be 0 and a 26.3433 percent chance that aaa_mod_var will be 1.
Node_1_99	If MAX_RATE is equal to 0 and COLLEXT_BAL is greater than or equal to 0 and is less than 88 and NUM_INQ is less than or equal to 1 then there is a 73.1132 percent chance that aaa_mod_var will be 0 and a 26.8868 percent chance that aaa_mod_var will be 1.

Translation

Structured Attribute

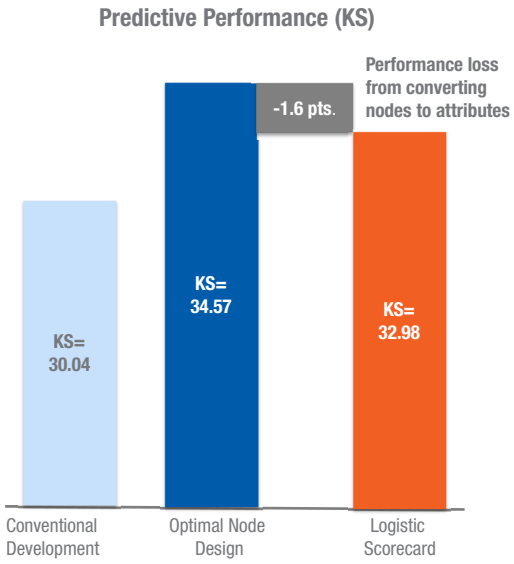
Medical Collection Accounts – Unpaid, >6 months age

		Number of inquiries		
Balance $	Age	1	2	3+
<=1,000	7-12 mos	32.6%		
	12-36 mos	30.4%		
	36+ mos	27.7%	28.7%	32.1%
>1,000	7-12 mos	32.7%	40.5%	
	12-36 mos	32.1%	34.6%	
	36+ mos	30.3%	34.5%	

Step 3: Scorecard Integration

While attribute conversion does cause some loss of performance, this has been minimized as much as possible. In the end, this process results in a significant performance opportunity, providing a final lift of more than 10 percent in the performance of the Dormant segment (Figure 7) and more than a 30 percent lift to the No Trade segment when compared with a traditional model. The final scorecards include multi-dimensional attributes designed for revolving products, installment products, inquiries and payment history, with the structured scorecards able to be aligned and fully integrated into traditional scoring algorithms.

Figure 7: Performance for the optimal node and the final logistic scorecard compared to conventionally developed scorecards [Holdout validation, 2014-2016]



CONCLUSION

As the financial industry works to expand the scoreable population in a responsible manner, credit score model developers also have a responsiblity to exhaust every possible avenue for accurate risk assessment of this 'credit invisible' population. Additional or alternative data sources clearly offer one such avenue; however, much needs to be done to bring these data to the same quality and coverage levels as mainstream credit data. As VantageScore 4.0 demonstrates, a guided application of machine learning techniques in model design may produce strong predictive insight from qualified, robust credit file data.

The VantageScore credit score models are sold and marketed only through individual licensing arrangements with the three major credit reporting companies (CRCs): Equifax, Experian and TransUnion. Lenders and other commercial entities interested in learning more about the VantageScore credit score models, including the latest VantageScore 4.0 credit score model, may contact one of the following CRCs listed for additional assistance:

EQUIFAX®

Call 1-888-202-4025

http://VantageScore.com/Equifax

experian.

Call 1-888-414-4025

http://VantageScore.com/Experian

TransUnion

Call 1-866-922-2100

http://VantageScore.com/TransUnion