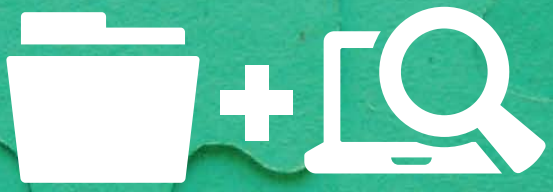


Modeling Blended Alternative and Traditional Data:

A Case Against Variable Reduction

White Paper

LexisNexis®
RISK SOLUTIONS

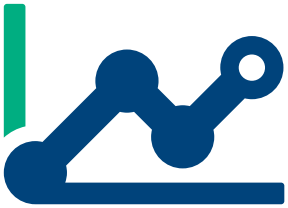
Credit scoring models that include alternative data have proven their predictive power in a wide variety of use cases across the financial services industry.

We generally define alternative data as differentiated and incremental to tradeline data reported by the three national credit bureaus.

Models driven by alternative data have a variety of use cases such as:

Full File Segmentation: Standalone scores that are often used alongside traditional scores like FICO® or VantageScore® to provide additional segmentation.

Thin-File and No-Hit Scoring: Standalone alternative data scores for consumers lacking sufficient credit bureau history to be traditionally scored.



Hybrid Modeling: Building scores that combine alternative data with other traditional credit tradeline data at the attribute level to create a more predictive solution.

Alternative data generally captures information not contained within traditional tradeline data sources. When alternative and traditional data is combined at the score or attribute level, the combination creates a broader information set that can be used to generate a more robust model. This combined set can fill in consumer behavioral gaps that may be exposed when considering only credit tradeline data.

The benefit of a combined dataset is obvious in a thin-file scenario where the consumer has little or no credit tradeline data. However, in the case of a consumer with a robust scoreable credit file, *it is easy to get the false perception that alternative data shows little to no lift over purely tradeline data.*

The following case study demonstrates how best to gain predictive value from alternative data in models that use traditional tradeline data.

Case Study: Sample and Variable Reduction

This example looks at a full file auto-lending portfolio with nationwide coverage.

The sample includes two snapshots:

The first is credit tradeline information at time of application for an auto loan

And the second is the performance of booked loans

As related to the sample, there were over 1,300 traditional and alternative data elements available as independent variables. The research model was created using forward stepwise logistic regression. Stepwise regression was used for the following reasons:

1. The model development process was entirely objective and based solely on the incremental predictive value of each variable.
2. The results were not subjected to the particular modeler biases with regard to variable selection.
3. It provided clear univariate feedback that would be more straightforward to interpret than if it were derived from more complicated non-linear methods.

Thirteen hundred variables can be cumbersome for modeling, compelling many data scientists to do some form of variable reduction in order to reduce the number of variables to explore. Variable reduction was conducted based on ranking the predictive power of the individual variable's correlation to the dependent variable and eliminating variables that do not perform over a certain threshold.

In this example, the alternative data fell outside of the top 250 variables with the highest-ranking variable landing at 296th. However, absolute rankings tell only part of the story. Alternative data is statistically independent of traditional tradeline data thus the incremental value of alternative data will be far greater than its standalone value.

We compared two models, one based on the top 250 variables (of which none were alternative data) and one based on all thirteen hundred plus variables. **Other than the variable reduction step, both models were created using the same overall techniques and subject to similar binning rules.** The only methodological difference is that the first model was based on the reduced list of 250 variables and the second allows any of the thirteen hundred plus variables to enter the model.

As would be expected, the first variable in each model was the same since both models started with the singularly most predictive variable. The models began to diverge starting with the second variable. The first model only had traditional data in the top 250 elements because it was limited to the top 250 most predictive variables, none of which were alternative data.

The second model chose an alternative data element as the second variable because it was the most uniquely predictive variable available for the stepwise algorithm. During the variable reduction step, this alternative data variable was ranked 296th on the predictive power list. As the forward regression continued, multiple other alternative data elements entered the model.

In total, alternative data elements made up roughly 25% of the model's overall predictive power even though none would have entered the model based on the top 250 predictive variables.

Here is the rank ordering table for both the variable reduced model (Table 1) and non-reduced model (Table 2).

Note that the overall KS value increases by 4.3% from the variable reduced model to the fully inclusive model.

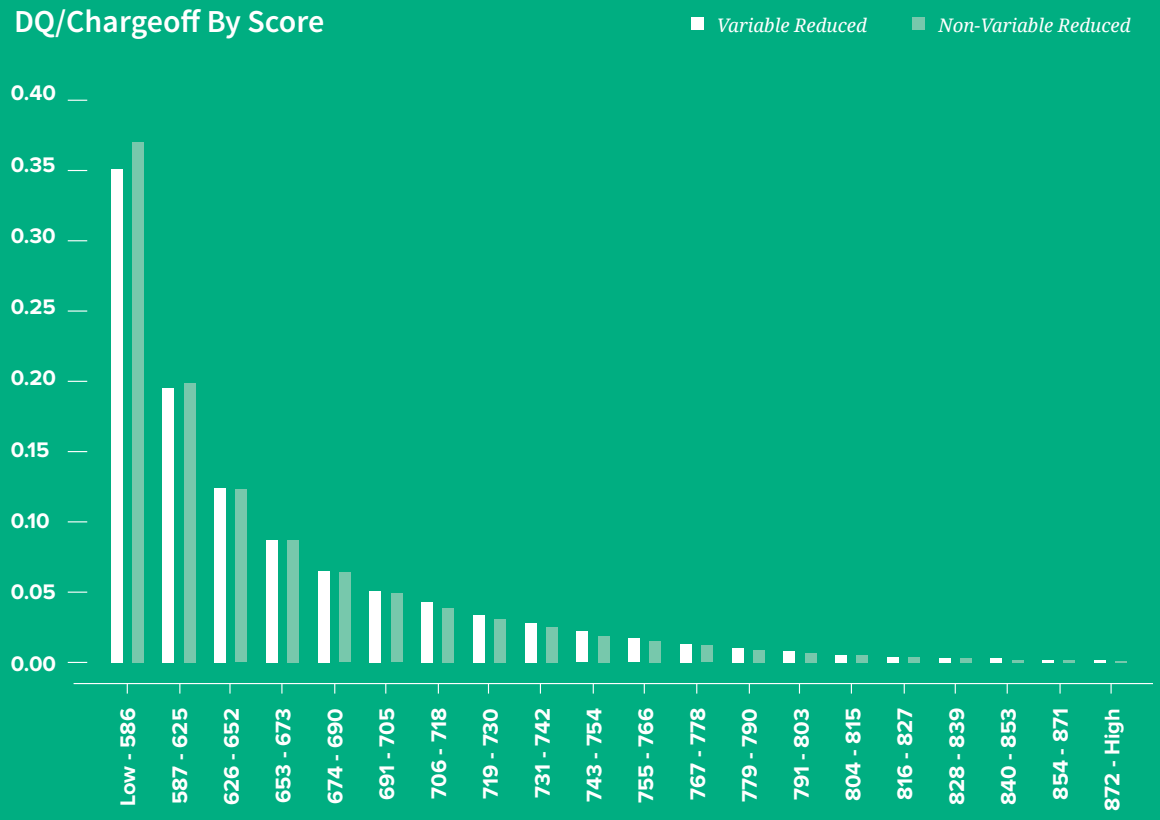
Table 1: Variable Reduced KS of 55

Score	# of Accounts	Cumulative % of File	# DQ/Chargeoff	% DQ/Chargeoff	Cumulative % of DQ/Chargeof
Low - 586	547,549	5.0%	191,950	**35.1%**	32.8%
587 - 624	556,619	10.1%	108,423	**19.5%**	51.4%
625 - 650	554,693	15.2%	68,630	**12.4%**	63.1%
651 - 669	547,033	20.2%	47,375	**8.7%**	71.2%
670 - 684	527,888	25.1%	34,302	**6.5%**	77.1%
685 - 699	564,163	30.3%	29,021	**5.1%**	82.0%
700 - 713	553,987	35.3%	23,554	**4.3%**	86.1%
714 - 726	533,172	40.2%	18,147	**3.4%**	89.2%
727 - 739	540,695	45.2%	15,009	**2.8%**	91.7%
740 - 752	549,065	50.2%	11,893	**2.2%**	93.8%
753 - 765	558,893	55.4%	9,241	**1.7%**	95.3%
766 - 777	526,324	60.2%	6,801	**1.3%**	96.5%
778 - 789	527,481	65.0%	5,352	**1.0%**	97.4%
790 - 801	547,397	70.1%	4,197	**0.8%**	98.1%
802 - 813	544,253	75.0%	2,942	**0.5%**	98.6%
814 - 824	553,195	80.1%	2,479	**0.4%**	99.1%
825 - 836	565,271	85.3%	1,968	**0.3%**	99.4%
837 - 847	532,053	90.2%	1,510	**0.3%**	99.7%
848 - 861	547,501	95.2%	1,201	**0.2%**	99.9%
862 - High	521,750	100.0%	831	**0.2%**	100.0%
	10,898,983		**584,827**	**5.4%**	

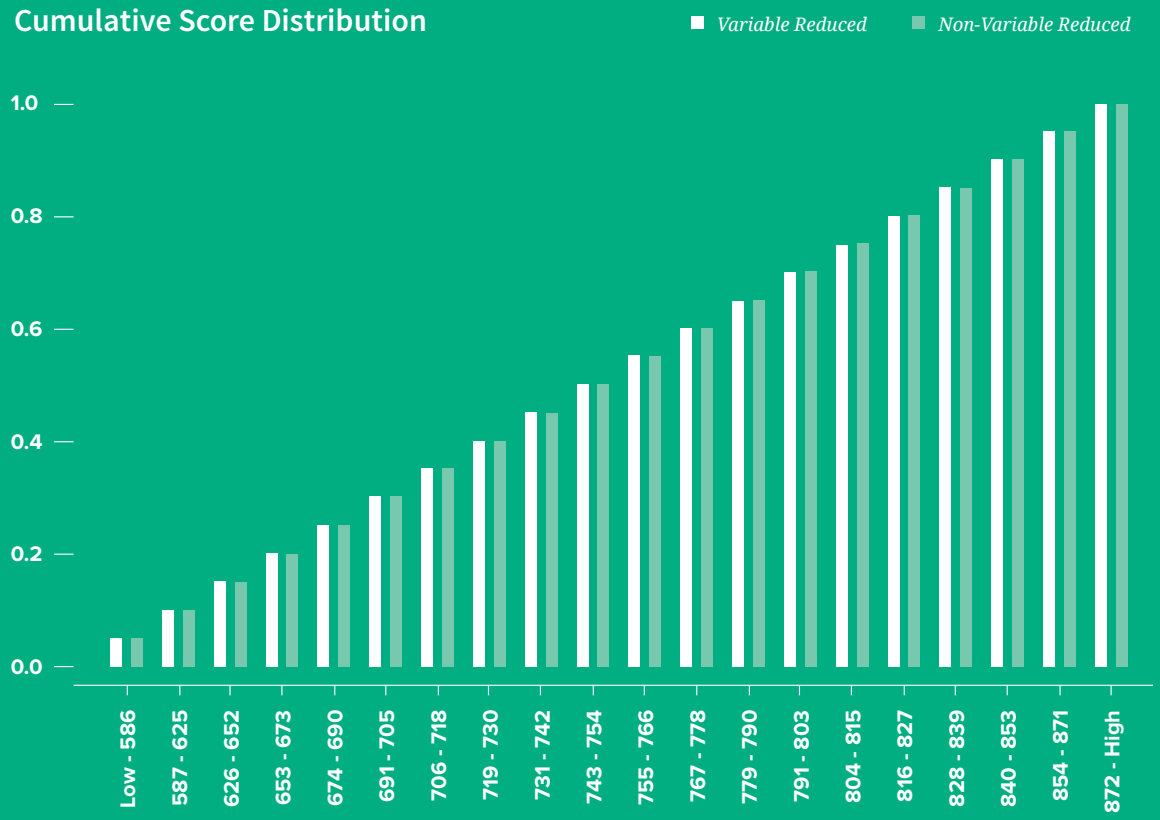
Table 2: No Variable Reduction
KS of 57.4

Score	# of Accounts	Cumulative % of File	# DQ/Chargeoff	% DQ/Chargeoff	Cumulative % of DQ/Chargeof
Low - 586	546,994	5.0%	203,889	**37.3%**	34.9%
587 - 625	555,123	10.1%	110,605	**19.9%**	53.8%
626 - 652	545,754	15.1%	67,302	**12.3%**	65.3%
653 - 673	547,124	20.1%	47,482	**8.7%**	73.4%
674 - 690	546,121	25.2%	35,056	**6.4%**	79.4%
691 - 705	562,971	30.3%	27,503	**4.9%**	84.1%
706 - 718	538,467	35.3%	20,774	**3.9%**	87.7%
719 - 730	526,986	40.1%	16,472	**3.1%**	90.5%
731 - 742	544,688	45.1%	13,555	**2.5%**	92.8%
743 - 754	551,686	50.2%	10,578	**1.9%**	94.6%
755 - 766	548,922	55.2%	8,030	**1.5%**	96.0%
767 - 778	547,431	60.2%	6,354	**1.2%**	97.1%
779 - 790	538,954	65.2%	4,834	**0.9%**	97.9%
791 - 803	572,154	70.4%	3,774	**0.7%**	98.5%
804 - 815	535,967	75.3%	2,675	**0.5%**	99.0%
816 - 827	546,391	80.3%	1,982	**0.4%**	99.3%
828 - 839	530,183	85.2%	1,440	**0.3%**	99.6%
840 - 853	553,407	90.3%	1,106	**0.2%**	99.8%
854 - 871	538,078	95.2%	920	**0.2%**	99.9%
872 - High	521,583	100.0%	496	**0.1%**	100.0%
	10,898,983		**584,827**	**5.4%**	

DQ/Chargeoff By Score



Cumulative Score Distribution



The results show the ability of the non-variable reduced model to more accurately score risk by moving poorly performing loans into the lowest four scoring tiers.

The bottom 5% of scores in the file show a 6.4% increase in the number of delinquent or charged off loans. This increase is composed of loans that migrated from higher score bands in the variable reduced model to a lower score band in the non-reduced model.

6.4%
INCREASE
IN THE NUMBER OF DELINQUENT OR CHARGED OFF LOANS IN THE BOTTOM 5% OF SCORES.

Notice also that the score distribution does not have a significant shift between the two models. This shows that the loans that moved were swapped out with performing loans that scored in the bottom tiers in the variable reduced model but higher in the score range in the non-reduced model. This increase in scoring accuracy allows the lender to add more precision in risk-based pricing the portfolio.

Further isolation of these poorly performing loans offers a competitive advantage to the lender over lenders that are using only tradeline data in their scores. *Increased confidence in a score's ability to identify risky loans allows a lender to trim their interest rate margin* ***without sacrificing profits.***

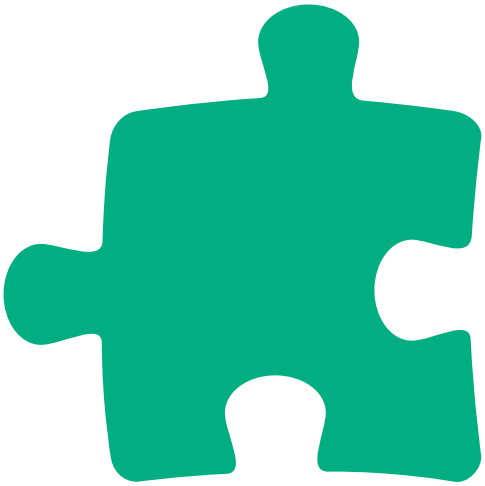
Conclusion

This case study is typical of the type of modeling paradigm we often see: While alternative data may not appear to add margin to traditional data when each variable is examined for its individual predictive power, a deeper look reveals the opposite.

Alternative data adds significant value when combined with credit tradeline data.

This demonstrative case study illustrates the need to dig deeper into alternative data in order to create a more robust and predictive solution. In this example, the model without variable reduction took a slightly longer time to produce, but clearly showed stronger results. **In a time-sensitive scenario, creating a model without variable reduction may not be possible.**

If this is the case, the best use of a modeler's time would be to **treat the data separately and perform variable reduction on both the traditional data and alternative data and then bring each independent variable set together to create a combined model.** By doing this and allowing the most powerful alternative data elements to combine with the traditional variables, a more robust model will be created.



Do you want to enhance the predictability of your credit scoring models with Alternative Data?

Contact us today at 800.869.0751 or visit lexisnexis.com/CreditRisk

LexisNexis®
RISK SOLUTIONS

About LexisNexis Risk Solutions

At LexisNexis Risk Solutions, we believe in the power of data and advanced analytics for better risk management. With over 40 years of expertise, we are the trusted data analytics provider for organizations seeking actionable insights to manage risks and improve results while upholding the highest standards for security and privacy. Headquartered in metro Atlanta, LexisNexis Risk Solutions serves customers in more than 100 countries and is part of RELX Group plc, a world-leading provider of information and analytics for professional and business customers across industries. For more information, please visit www.lexisnexisrisk.com.