MIS 749: Business Analytics San Diego State University

FICO Analytic Cloud

Credit Risk and Interest Revenue Assessment

Steven Felger | Julie Errickson | Roger Chen | Samuel Beckom

[Date]

Executive Summary

As part of a new collaboration with Fair Isaac Corporation (FICO) and San Diego State University this group of students had the opportunity to test and employ FICO's new Analytic Cloud software in order to address a given business case. The case presented a challenge to support a credit grantor for mortgage and home equity line of credit (HELOC) regarding its credit risk and interest revenue. The given data for the case represents three years of information collected from applicants regarding terms of a loan, FICO score, credit bureau information and other additional attributes. The wealth of the data provided allowed the team to use two software products to develop a model and decision strategy: Analytic Modeler Scorecard Professional and Analytic Modeler Decision Tree Professional. In leveraging this advanced user-friendly software, the team was able to assess the data and provide the credit grantor a model that greatly improves upon their current model in order to enhance the process to book accounts that will pay as negotiated and return the highest revenue.

I. Overview and Case Introduction

The FICO Credit Risk and Interest Revenue case provided student analysts the challenge and opportunity to improve upon a credit grantor's current home equity line of credit (HELOC) approval model. The situation at hand called for the group to assess the efficacy of the current screening process for loan approvals. We performed data analytics with a given data set that provided three years of loan application information collected from applicants and their credit bureau account history, including attributes such as FICO score, trade line statistical information, and other attributes regarding the borrower's ability to pay, value of the property to be used as collateral, and the loan amount requested. In addition to these variables, we were given the dependent variables 'paid-as-negotiated' and 'interest revenue collected to date' to assess the success of the screening process approvals. The data set did not contain information regarding applicants who did not pass the screen, nor applicants who passed the screen but decided not to take the loan. The case provided a great opportunity to employ data mining from a process perspective in a supervised modeling environment and utilized prediction methodologies such as linear regression and regression trees as well the classification methodologies logistic regression and classification trees. The core ideas focused on in developing the model for this project were logistic regression in classification (making a prediction about a future event by assigning it to one of two or more classes) and linear regression in prediction (trying to estimate the value of some future event).

II. Process and Data Approach

Following the steps for data mining, as presented in MIS 749, we had a unique challenge in determining our task because we had both a classification (paid-as-negotiated) dependent variable and a prediction (interest revenue) dependent variable. Luckily, as we began to develop our understanding of the project and explored the data, the analytic cloud software made the visualization of the data and complexity of having two distinct dependent variables very easy to comprehend. Had this instead been done using Microsoft Excel, there were have been a much steeper challenge with time and complexity in organizing the data and determining our starting point. Another factor we benefited from was that the given data was already cleaned and preprocessed. These factors and support from FICO allowed the group to get right into separating the data.

After gathering a good understanding of the task at hand and getting acclimated with the data and understanding our 33 variables, we partitioned our training data and validation by setting a filter. In class we were taught to use a 60% training data and 40% validation data, but because of the cleanliness of the data, we chose to go with FICO's standard of 70% training data and 30% validation. In order to group our data we had many options to transform the variables to get a linear relationship, but due to the steep learning curve with the software and for the sake of limited team capacity, we chose to forego this option and utilize the software's classification feature and developed our 'bins' in a monotonically increasing fashion. This reduced the time and cost involved with manipulating the data so that we could focus on selecting the right variables for inclusion in the model.

Keeping in mind that we were tasked to improve upon the current model, we needed to set the baseline impact and efficacy of the current custom score (given model). We built the current custom score model in Scorecard Pro and identified an R^2 of 0.164 and an ROC of 0.759. Following that, we decided to build the original model, consisting of the following variables and rules: Loan-to-value (LTV) < 100, Debt-to-income ratio (DR) < 50, FICO score > 600, and a

current custom score > 200. This resulted in an R^2 of 0.203 and an ROC of 0.791. This original model set the target for us to beat with our recommendations for improvement.

Once this base lining process was complete, we began to try numerous variables in concert to develop a bountiful array of models. With some loan industry expertise on the team we began to apply reason and logic in our variable selection using industry standards like 80% or less LTV and Debt to Income of 40% or less, but quickly found this was not as beneficial as we had hoped in deriving a more effective model. This may go to show that following the primary mortgage loan standards and what is widely accepted may not yield the best results for a specific credit grantor of HELOC loans. So we quickly abandoned this approach and began with a data-driven approach with the plan to assess a few competing models by applying domain expertise post hoc. The software provided a great feature that allowed us to guess and check and constantly improve upon our model with 'what-if' scenarios. It also allowed us to see the independent R^2 contribution for each variable by itself, but also provided the marginal contribution to R^2 based on the current variables already selected. We found this to be one of the most beneficial aspects of the software. After trying countless models, we were able to identify our best model and balance of variables to include inquiries, delinquency, and utilization. Once we completed our new model, this was transferred over to Decision Tree Pro to be manipulated even further in developing the final optimal strategy to complement our optimized model. But first, we'll introduce the data and representative statistics that supports our confidence in our final model.

III. Scorecard and Model Statistics

Before we run our own models, we need to know what we will benchmark against. The previously mentioned current custom score was not the only model we needed to benchmark against, we also benchmarked against the current complete custom score. Therefore, we ran both of the models through the tool, and picked the current complete custom score to continue the discussion because that was the stronger and better model with which to make a comparison.

LR^2: 0.203 • **D:** 1.155 • **K-S:** 0.455 *(@ 3.844)* • **ROC:** 0.791 • **Gini:** 0.583

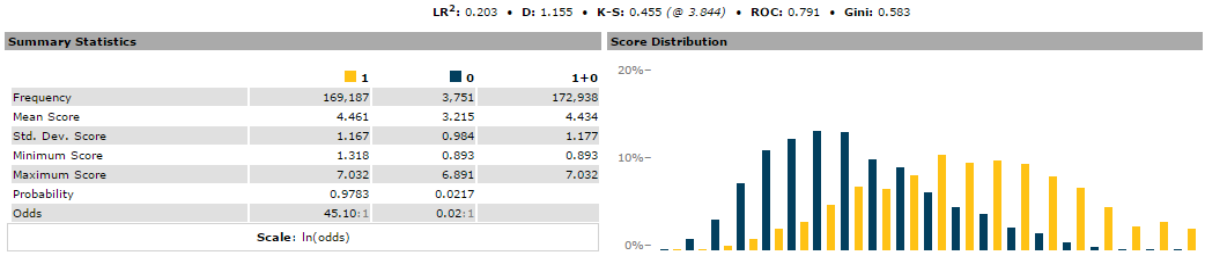
Rank	Predictor	Independent LR^2	Contribution
1	Current_Custom_Score	0.164	0.059
2	FICO_Score	0.152	0.031
3	LTV	0.038	0.018
4	Debt_Ratio	0.007	0.007
Step Two			
5	Income	0.025	0.014
6	Collateral_Value	0.031	0.014
7	Bank_Relationship	0.037	0.012
8	Nb_Borrowers	0.023	0.009
9	Loan_Amt_Req	0.038	0.008
10	CB_Nb_Sat_TL	0.018	0.008
11	CB_Nb_Total_TL	0.011	0.007
12	Prior_Custom_Score	0.076	0.005
13	CB_Rev_Util	0.094	0.003

Figure 1. Current Complete Custom Score Summary Statistics

To proceed with an in-depth explanation of the current complete custom score, it is necessary to discuss the model statistics from Figure 1. Logistic R^2 (LR^2) is a model performance measure for binary-performance models. It is analogous to the R^2 model performance measure for continuous-performance models.

Essentially, $LR^2$of 0.203, D of 1.155, and ROC of 0.791 are the benchmarks that our model will be compare against. In the current complete custom score model, the variables ranked 1 - 4 are the step one variables, which are the variables that run the model. The rest of variables

are classified as step two variables, which are the second best candidate variables. Independent LR^2 is the independent contribution to the model from a specific variable. Contribution is the estimated decrease of LR^2 realized when removing an independent variable from step one of the model. It's also necessary to look at the score distribution from the model in Figure 2.



	1	0	1+0
Frequency	169,187	3,751	172,938
Mean Score	4.461	3.215	4.434
Std. Dev. Score	1.167	0.984	1.177
Minimum Score	1.318	0.893	0.893
Maximum Score	7.032	6.891	7.032
Probability	0.9783	0.0217	
Odds	45.10:1	0.02:1	

Scale: ln(odds)

Figure 2. Current Complete Custom Score Distribution

Note that the 1s with the yellow bars are the individuals who did pay the loan as negotiated, and the 0s with the blue bars are the individuals who defaulted. Later on when we determined the loan approval cutoff line on the bar histogram, everyone on the right side of cutoff line would ideally be approved for the loan and everyone else on the left side would be denied a loan. Therefore, a perfect score distribution would have two sets of bars completely separated from each other without any overlap.

Before building up our own model, we also needed to group the bin breaks for each variable as indicated in Figure 3 and Figure 4.

Predictor: Age (Notes) | Perf: Risk_Flag

C	A	Level	Select	Merge	<	>	Neu.	#Raw	#1	#0	W%1	W%0	W%	P(1)	WoE	IV	WoE
8		**Total**							**3,533**	**3,817**	**100.0**	**100.0**	**100.0**	**0.978**		**0.101**	**0.101**
0		-Inf -< 29				↓		10	100	260	2.8	6.8	2.9	0.949	-0.878	0.035	
1		29 -< 32			↑	↓		3	172	261	4.9	6.8	4.9	0.970	-0.340	0.007	
2		32 -< 33			↑	↓		1	91	85	2.6	2.2	2.6	0.981	0.146	0.001	
3		33 -< 38			↑	↓		5	404	610	11.4	16.0	11.5	0.970	-0.335	0.015	
4		38 -< 54			↑	↓		16	1,636	1,727	46.3	45.2	46.3	0.979	0.023	0.000	
5		54 -< 80			↑	↓		26	1,093	802	30.9	21.0	30.7	0.985	0.387	0.038	
6		80 - Inf			↑			14	37	72	1.0	1.9	1.1	0.961	-0.588	0.005	
7		Other					•	0	0	0	0.0	0.0	0.0	0.000	0.000	0.000	

Figure 3. Before Variable Bin Breaks Grouping

Predictor: Age (Notes) | Perf: Risk_Flag

C	A	Level	Select	Merge	<	>	Neu.	#Raw	#1	#0	W%1	W%0	W%	P(1)	WoE	IV	WoE
6		**Total**							**3,533**	**3,817**	**100.0**	**100.0**	**100.0**	**0.978**		**0.083**	**0.083**
0		-Inf -< 29				↓		10	100	260	2.8	6.8	2.9	0.949	-0.878	0.035	
1		29 -< 32			↑	↓		3	172	261	4.9	6.8	4.9	0.970	-0.340	0.007	
2		32 -< 38			↑	↓		6	495	695	14.0	18.2	14.1	0.972	-0.262	0.011	
3		38 -< 54			↑	↓		16	1,636	1,727	46.3	45.2	46.3	0.979	0.023	0.000	
4		54 - Inf			↑			40	1,130	874	32.0	22.9	31.8	0.984	0.334	0.030	
5		Other					•	0	0	0	0.0	0.0	0.0	0.000	0.000	0.000	

Figure 4. After Variable Bin Breaks Grouping

Figure 3 and Figure 4 show the relationship between a specific predictor and the performance variable. In this case, the variable Age contains several bin breaks with separated

pieces. It was our job to regroup them so that we could identify a clear relationship between Age and Risk. Weight of Evidence (WoE) is a log transformation of weighted percentage of 1 (W%1) divided by weighted percentage of 0 (W%0), which is also translated to the horizontal bar histogram in the far right column of the chart. The bar to the left with negative spacing means that the specific bin group is more populated by 0s. Vice versa, the bar to the right with positive spacing means that the specific bin group is more populated by 1s. It was difficult to determine a compelling relationship between Age and Risk before the grouping. As an example, since there are only people of age 32 or 33 from group #2, we would have a really hard time buying any information from that specific age group. Thus, a merge of groups 2 and 3, and groups 5 and 6 allowed us to get a clean shape on WoE as well as getting rid of the red bar, which indicates insufficient number of data. Additionally, we can conclude that as age increases, 1 is more populated than 0. In other words, people are more likely to pay the loan as age increases. After performing the same type of merging and grouping for all 33 variables, we were ready to build our first iteration of model as indicated in Figure 5.

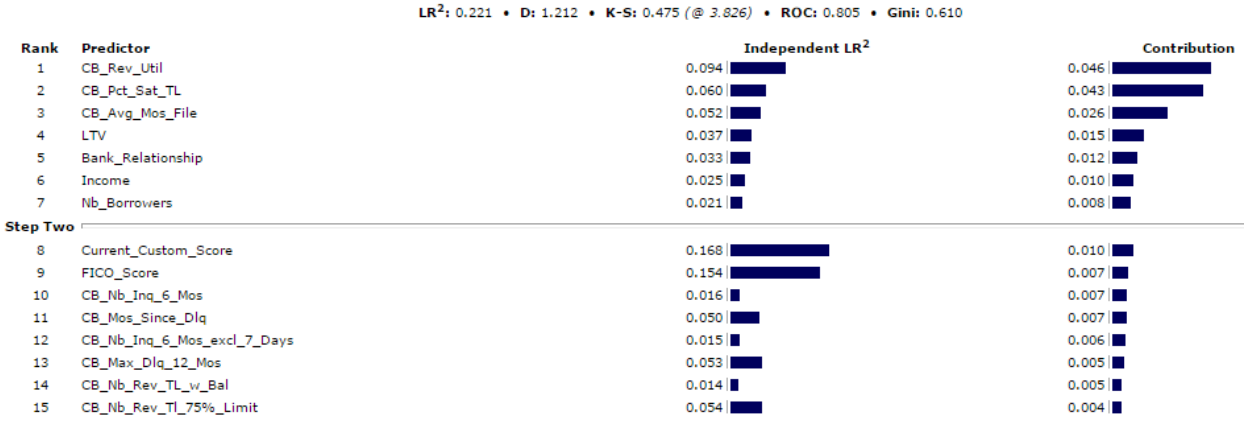


Figure 5. First Iteration Model Statistics

After the first attempt of building the model, a higher LR^2 of 0.221 was derived compared to the current complete custom's LR^2 of 0.203 as well as a higher ROC of 0.805 from 0.791, which was a fantastic model to start with. To take further consideration of the business nature of lending and to take inquiries, delinquency, and utilization into account, another round of variables were selected for the second iteration, which is presented in Figure 6.

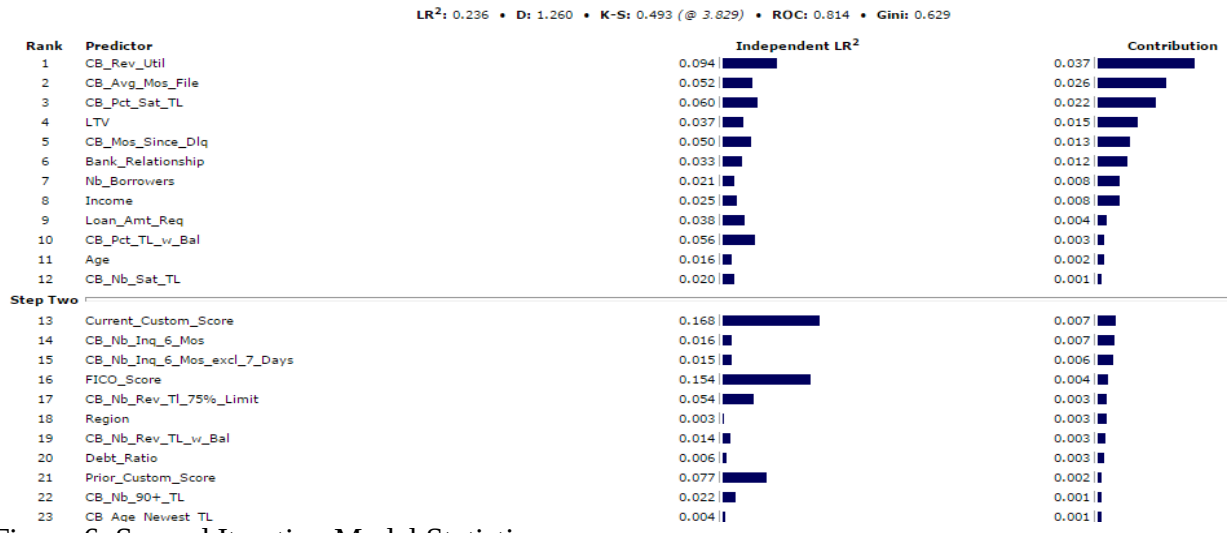


Figure 6. Second Iteration Model Statistics

The LR^2 of 0.236 and ROC of 0.814 represent another substantial improvement from the first iteration. However, another problem was realized when some step-two variables performed better than the step-one variables in this model from contribution. So the next step involved moving some top performing step-two variables with higher contribution into the step one group. After playing around with program, we were able to develop our final model with the best indicator statistics as shown in Figure 7.

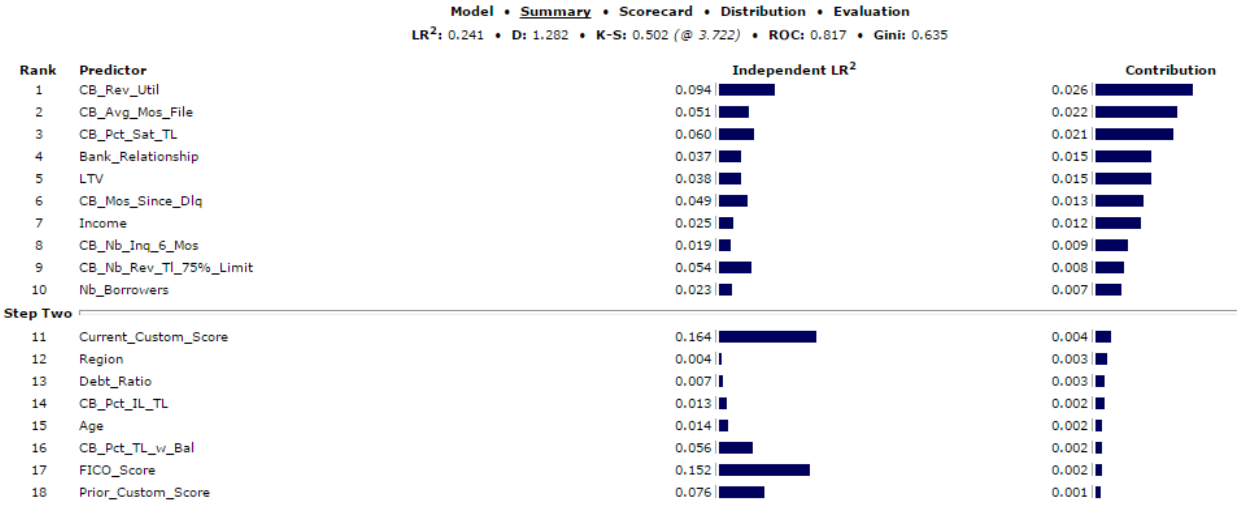
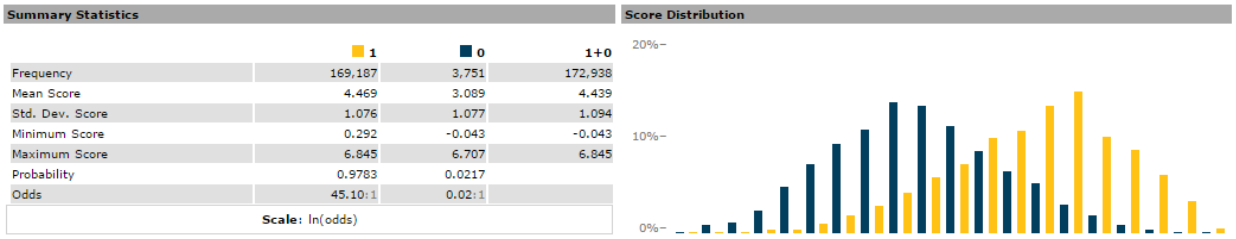


Figure 7. Final Best Model Statistics

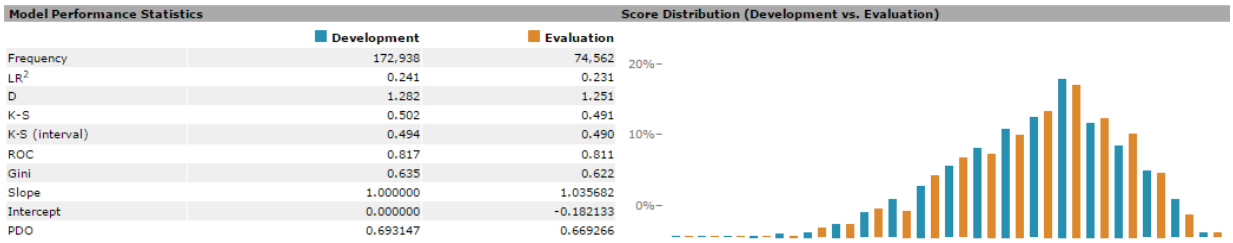
With the best LR^2 of 0.241 and ROC of 0.817, along with a perfect contribution distribution across the step-one and step-two variables, a superior final model was developed. A further analysis was continued for the score distribution in Figure 8.



Summary Statistics	1	0	1+0
Frequency	169,187	3,751	172,938
Mean Score	4.469	3.089	4.439
Std. Dev. Score	1.076	1.077	1.094
Minimum Score	0.292	-0.043	-0.043
Maximum Score	6.845	6.707	6.845
Probability	0.9783	0.0217	
Odds	45.10:1	0.02:1	
Scale: ln(odds)			

Figure 8. Final Model Score Distribution

A beautifully distributed skewness of 1s and 0s in the score distribution indicates this model is very neat and efficient, which is further confirmed with a D score of 1.282 compared to 1.155 and a K-S score of 0.502 compared to 0.455 from the current complete custom score. Next the model was tested using the training data.



Model Performance Statistics	Development	Evaluation
Frequency	172,938	74,562
LR^2	0.241	0.231
D	1.282	1.251
K-S	0.502	0.491
K-S (interval)	0.494	0.490
ROC	0.817	0.811
Gini	0.635	0.622
Slope	1.000000	1.035682
Intercept	0.000000	-0.182133
PDO	0.693147	0.669266

Figure 9. Final Model Score Distribution Development vs. Validation

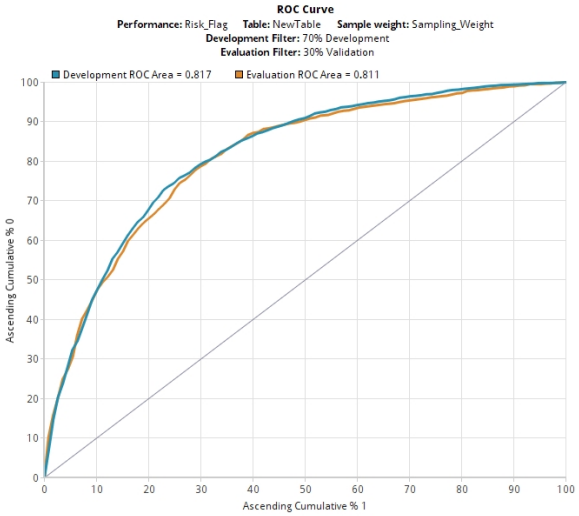


Figure 10. Final Model ROC Curve Development vs. Validation

From Figure 9, the development bars and validation bars closely match each other. From Figure 10, the two ROC curves overlap each other closely. Both figures tell the same story, that the final model is indeed tremendous and efficient. Now we will compare our final model to the current custom model and the current complete custom model shoulder to shoulder with each ROC curve and lift curve from Figures 11 and 12.

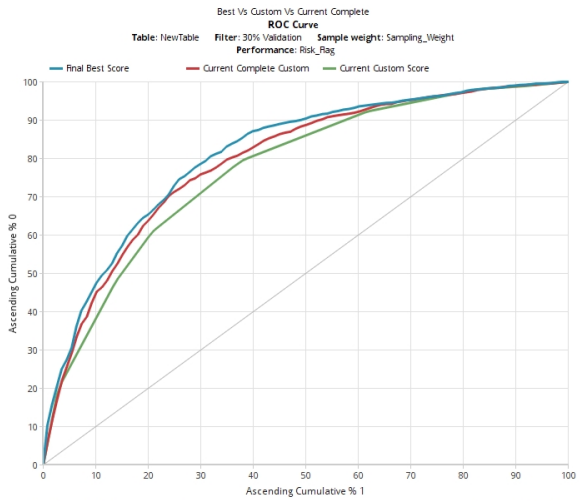


Figure 11. ROC Curve Final Model vs. Current Custom Model vs. Current Complete Custom Model

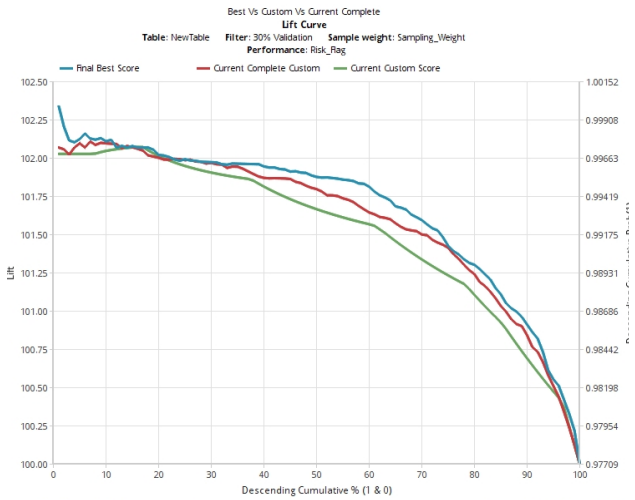


Figure 12. Lift Curve Final Model vs. Current Complete Custom Model vs. Current Custom Model

As expected, our final model, which is the blue line, outperforms both the current custom model and the current complete custom model. The next section of this paper provides further verification and evaluation based on some domain of expertise of the industry.

IV. HELOC Financial Industry Perspective

A loan by definition is the purchase of the present value use of money, combined with the promise to repay that money in the future according to an agreed upon schedule and at a specified rate of interest, which may also be adjusted periodically in relation to an agreed upon published benchmark rate. A HELOC or Home Equity Line Of Credit is a loan which is secured by the collateral of a specifically named real estate property. A HELOC is similar to a conventional primary mortgage in that the interest expense is tax deductible, but instead of being a fixed payment installment loan, this is a revolving loan similar to a credit card, which means the minimum payment amount will vary with the size of the outstanding balance ("Home Equity Line of Credit.").

This type of loan is considered to be a second mortgage by definition, with more inherent risk than a primary mortgage. When a mortgage lien has second position in payout priority as a creditor behind the primary home loan lender, there is a higher risk of loss to the second position lender in the event of a default by the borrower. Since the primary mortgage loan is repaid first, only the remaining proceeds of a foreclosure sale left over after repaying the outstanding principal of the primary mortgage is available to repay the outstanding principal of the second position loan, which is the HELOC loan we are modeling. When a primary mortgage is refinanced, the second position mortgage lender is typically required to subordinate its claim on the real estate collateral to maintain the relative positions of primary and secondary loans, as the payout priority in the case of default is done based upon the earliest recording date, unless an agreement exists to change that priority, which is typically a refinance requirement. A HELOC loan therefore has more risk than a primary mortgage in the event of a default, and it should be priced accordingly through the interest rate charged to the borrower. Careful consideration

should be given to determine the borrower's overall repayment ability before granting credit to a borrower or set of borrowers for a HELOC loan ("Second Mortgage").

The interest rate of a HELOC loan is typically variable, and is often higher than the rate on the primary mortgage to compensate for the additional risk of being in second position. A HELOC loan is typically paid out in smaller sums as needed to the borrower up to a certain fixed amount limit set by the lender. Since it is secured by a lien on real property, a HELOC loan has less default risk in general than an unsecured loan like a credit card, so the rate paid for a HELOC loan is often higher than a primary mortgage, but lower than a credit card with all other variables constant. HELOC loans are generally granted for a specific purpose, such as a home improvement project like the replacement of a roof, a remodel, room addition, or the construction of a pool. Often these loans are used to improve the property value of the associated real estate, which would increase the collateral value that secures the lien, thus also increasing the surplus home value available in the event of a default.

Economic conditions also need to be considered when granting a HELOC loan to a borrower, as the value of the underlying property can depreciate in a recessionary environment that is characterized by higher unemployment and lower wages, reducing the demand for homes and thus also reducing the average price paid by a willing buyer to a willing seller. When a loan amount matches the value of the home, there is zero equity in the property, the buyer's share of the property value above the bank's share if the house was to be sold at that point in time. A common measure used by banks to recognize this fraction of the loan amount over the current value of the underlying real estate property is called the Loan to Value ratio. It is a critical measure in determining the amount of ongoing risk to the bank, as it indicates the residual property value the borrower would be walking away from if the borrower defaulted on the loan.

A property with a 100% or greater loan to value ratio means the buyer is paying more than the fair value of the property to the bank at that moment. Therefore, a borrower with a 100% or greater LTV has a high likelihood of giving that house back to the bank without making any further payments if the expected future home price trend is neutral or declining, since the home has no positive residual value for the borrower if it was to be sold at that point in time. Applicants in our Credit Risk and Interest Revenue Assessment scenario are required to have a loan to value ratio of less than 100, and that assumes the current loan offering is included in that ratio or the borrower does not qualify for a HELOC loan.

The debt-to-income ratio is another significant measure of a borrower's ability to make regular payments on a loan. The ratio is calculated using the borrower's total current debt payments over the amount of regular income that is received during that same period, such as per month or per year. Applicants in this scenario are required to have a debt-to-income ratio of less than 50, and that assumes inclusion of the payment amount for the HELOC loan ("Debt-To-Income).

The FICO score is a highly accepted model used to evaluate a borrower's ability to pay back installment and revolving debt payment loans, like home, car, and student loans, property rentals, credit card debt, and other services like utility payments. FICO scores are generally broken into ranges which are used to separate reliable repayment history from inconsistent repayment history and no history. In general, the FICO score range of 300 to 500 is regarded as Bad Credit, which has a high likelihood of default. The next range of 501 to 600 is considered to be Subprime or Poor. These borrowers have a history of difficulty maintaining timely payments. Acceptable or Fair credit history is generally bracketed in the 601 to 660 FICO score range. Good credit is considered to be in the range of 661 to 780, while Excellent credit brackets the

FICO score range of 781 to 850, and these highest quality borrowers will experience the lowest rates a lender has available. In this scenario, applicants must have a minimum FICO score of 600, Good to Excellent Credit, to apply for a HELOC loan (Detweiler, G.).

The original Current Complete Custom Score model we were asked to evaluate in this assessment of a better home equity line of credit granting methodology used four variables: Current_Custom_Score, FICO_Score, LTV, and Debt_Ratio. Please refer back to Figure 1, Current Complete Custom Score Summary Statistics.

The Current_Custom_Score is the most significant variable in the existing model. It is a proprietary custom score, similar to the FICO score, used to determine a potential borrower's likelihood of repaying a mortgage or HELOC loan. This score was created by using an undisclosed mix of inputs that are available in part on a customer's consumer credit report history, and some inputs may also be collected during the loan application process. This variable was given a minimum score of 200 to be included as an applicant in the data.

The FICO_Score is the second most significant variable in the existing model. This is also a proprietary custom score for determining a potential borrower's likelihood of repaying which also relies on information in a borrower's credit report history. This score however, is commercially available and has become a well-known lending industry standard, commonly used in determining lending rates to be offered to a borrower. A FICO score is comprised with 35% of a borrower's payment history, 30% of a borrower's credit utilization or percentage of available credit that has been used, 15% uses the length of the credit history, 10% has to do with how much of the credit is new, and 10% is scored based on the credit mix, such as the ratio of revolving and installment loans. (Simon, J.)

The third most significant variable in the original loan model is the LTV, or Loan-To-Value percentage of a real estate property. This ratio is the fraction of the desired loan amount over the appraiser assessed resale value of the home at that point in time of the application. This ratio tells the lender how much equity the borrower would have in the property if the requested loan is granted. The industry standard is generally an 80% limit for the loan to value ratio in a normal economy, however in a declining value market this ratio may be capped at 70%, while in a rapidly rising value market this ratio has been raised to 90% and higher.

The fourth variable used in the original model is the Debt_Ratio. The debt-to-income ratio is a measure of a person's ability to pay a loan obligation based on the ratio of total debt obligations over gross income. This is generally a two part ratio. The front end ratio is the percentage of income going to housing expense like rent or mortgage debt payments, including principal, interest, insurance and taxes. The back end ratio is the total of all debt payments, including housing, credit card minimum payments, auto loan, student loan, child support, alimony, and legal judgement payments. The debt to income ratio has limits that vary as the conventional loans are typically 28/43, while FHA loans are between 31/43 and 33/45, while VA loans are a flat 41% debt to income ratio. This scenario has capped the applicant debt to income ratio at 50, which then would imply that in this case we are assuming this is the total of all regular minimum debt payments including housing expense and other minimum debt payments like credit cards over regular income for a specified period, such as per month ("Debt-To-Income Ratio").

Through an iterative process, the final model of the following 10 variables from the original 33 available is shown in Figure 13. This model was chosen due to it having the highest Independent LR^2 of the iterative models at 0.241, with no significant contribution remaining

from the other 23 variables that were not selected, as the LR^2 would be diminished with any modifications beyond this final scorecard model.

Rank	Predictor	Independent LR^2	Contribution
1	CB_Rev_Util	0.094	0.026
2	CB_Avg_Mos_File	0.051	0.022
3	CB_Pct_Sat_TL	0.060	0.021
4	Bank_Relationship	0.037	0.015
5	LTV	0.038	0.015
6	CB_Mos_Since_Dlq	0.049	0.013
7	Income	0.025	0.012
8	CB_Nb_Inq_6_Mos	0.019	0.009
9	CB_Nb_Rev_Tl_75%_Limit	0.054	0.008
10	Nb_Borrowers	0.023	0.007

Figure 13. Final Scorecard Model – Step 1 Variables

The first most significant factor with the highest contribution score chosen in predicting the outcomes of the Risk_Flag and Interest_Revenue variables is the variable CB_Rev_Util, described as the Net Fraction Revolving Burden. This variable indicates the percentage of the borrower's total credit bureau reported limit that is currently used. This highly predictive indicator is preferable when it has a lower ratio of used credit over available credit. A good ratio for this variable is usage of 10% or less of the total revolving credit card available balance. When a borrower has established credit available and when that credit is only being minimally utilized, there is a higher likelihood that a borrower will be able to consistently repay his or her debts in a timely manner, particularly when there is a significant amount of available credit and that person is not heavily burdened with short term revolving debt payments ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The second most valuable predictor is the CB_Avg_Mos_File variable, described as the Average Months in File. This is the average of the calculation of the life in months of each credit account on file from inception until the present or until the date of closure. The longer the average life of the active credit history is the better, as it gives a long term picture of the borrower's payment history. This variable is similar in some ways to the variable Age, but predicts payment ability more specifically based on the average life of the credit history. Having no credit history at all is a detriment, as there is no record of payment behavior, good or bad ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The third variable in significance is the CB_Pct_Sat_TL, known as the Percent Trades Never Delinquent. This variable takes the ratio of the number of credit accounts that have never been delinquent over the total number of credit accounts. It is a significant indicator in identifying the borrowers who have accounts that never had any delinquency, which is clearly superior over an indicator of how many months of delinquency a borrower has had, as it is ranking a higher quality borrower with a perfect or nearly perfect credit history, as compared to a variable which measures how delinquent a borrower has been, categorizing instead a lower quality borrower ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The fourth variable in contribution importance to our model is the Bank_Relationship, which is also described as the Prior or Current Bank Relationship a borrower has had with our subject bank. This indicator has the possible values of New Customer, Current Mortgage, Prior

Mortgage, Current Deposit, and Prior Deposit. A banking customer who has an active current mortgage with our subject bank is our best choice for offering a HELOC loan to, assuming that is a primary mortgage. We would in that case have payment history in the bank records, and an existing valuable relationship with that person. Having a customer with savings or checking deposits with the subject bank is also a plus, as we can see they have assets on hand that can be used to make loan payments. A customer with a prior mortgage would be next, then current deposit, prior deposit, and lastly a new customer. A customer who has demonstrated ability to pay a mortgage is very likely to pay in a timely manner. In addition, the existing customer relationship generally indicates that this customer is happy with the bank's services and will likely continue to perform as in the past.

The fifth variable in contribution importance is the LTV or Loan-to-Value Ratio, which we have already discussed at length as being a very important indicator of the value of the underlying collateral in relation to the amount of the total loans held against the property, including the value of the proposed HELOC loan. Typically the LTV is capped at 80% on a primary mortgage, but there tends to be more variety in the maximum LTV offered with a HELOC loan since these loans are typically smaller and often are given for home improvement purposes. The maximum LTV allowed in the application data is 100%.

The sixth variable in contribution importance is the CB_Mos_Since_Dlq or Months since Most Recent Delinquency. This variable reports the number of months since the last reported delinquency on the accounts included in the borrower's credit report. The greater the number of months since delinquency the better, and there is a cap of 999 months which indicates no delinquency. This is another great indicator which focuses on selection of the highest quality borrowers ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The seventh variable in contribution importance to our model is Income, or the Applicant Total Income. This variable could be reporting the income of a single borrower, or the income of two or more borrowers if there is a spouse or other co-borrower. Income is a very important indicator as it indicates a borrower's ability to repay a loan that is specific to the size of the loan amount requested. This variable would have a high correlation with the debt to income ratio, however, this variable was superior to the debt to income variable in determining which borrowers were more likely to pay as negotiated.

The eighth variable in contribution importance is CB_Nb_Inq_6_Mos or the Number of Inquiries in the Last 6 Months. This variable tells us the number of inquiries recorded in an applicant's credit history in the past six months. This is a significant variable as it indicates if the borrower has been shopping for new credit due to a financial hardship or it could also indicate that other lenders have been turning this borrower down if there are a significant amount of inquiries but no or few new accounts. Unsolicited offers by lenders for marketing purposes are not included in this number ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The ninth variable in contribution importance is CB_Nb_Rev_Tl_75%_Limit or Number Bank/Natl Trades with a Balance of 75pct or Higher. This variable indicates the borrowers who have revolving credit lines that are at or above 75% of the high balance, which is a substituted variable for the unreported credit limit. This variable is very similar to the Net Fraction Revolving Burden, which was our first variable, however, it removes some of the smoothing done by the average calculation of the first variable's revolving credit lines. This variable acts somewhat more like a categorical variable by indicating how many credit lines are at or near full

usage, which could be an indicator of a higher risk, in that a borrower is maximizing the lower interest rate lines of credit, and perhaps has some credit lines with much higher interest rates that could be a problem if they are used when times get tough ("Credit Card," "Credit History," "Equifax," "Experian," "Key Dimensions," "Transunion").

The tenth or final variable in our model is Nb_Borrowers or Number of Borrowers. This variable is significant in that having more than one borrower significantly reduces the risk that the account will be paid late. Having two sources of income reduces the payment burden in times of job loss. The likelihood of both borrowers losing their jobs at the same time is much less than when you have only a single borrower. This last variable had a contribution score of 0.007, which is the lowest contribution margin of the variables we accepted to be used in our final model. The remaining variables in Figure 14 were each considered for the final model, however the LR^2 of the model was reduced when any of these variables was added in as a predictor. Therefore, the model had reached peak performance with the ten variables shown in Figure 13.

Rank	Predictor	Independent LR^2	Contribution
Step Two			
11	Current_Custom_Score	0.164	0.004
12	Region	0.004	0.003
13	Debt_Ratio	0.007	0.003
14	CB_Pct_IL_TL	0.013	0.002
15	Age	0.014	0.002
16	CB_Pct_TL_w_Bal	0.056	0.002
17	FICO_Score	0.152	0.002
18	Prior_Custom_Score	0.076	0.001
19	CB_Nb_Sat_TL	0.018	0.001
20	Loan_Amt_Req	0.038	0.001
21	CB_Nb_Total_TL	0.011	0.001
22	CB_Nb_Rev_TL_w_Bal	0.016	0.001
23	CB_Nb_90+_TL	0.021	0.001
24	CB_Max_Dlq_12_Mos	0.041	0.001
25	CB_Age_Oldest_TL	0.037	0.001
26	CB_IL_Util	0.004	0.001
27	Collateral_Value	0.031	0.001
28	CB_Age_Newest_TL	0.004	0.001
29	CB_Nb_IL_TL_w_Bal	0.002	0.001
30	CB_Max_Dlq_Ever	0.039	0.000
31	CB_Nb_60+_TL	0.029	0.000
32	CB_Nb_TL_Open_12	0.004	0.000
33	CB_Nb_Inq_6_Mos_excl_7_Days	0.017	0.000

Figure 14. Final Scorecard Model – Step Two Variables

The ten variables used in our analysis proved to be the best combination of the 33 options available in predicting which borrowers would be most likely to pay as negotiated, and which borrowers would best help us to maximize our interest revenue. The remaining 23 variables, as seen in Figure 14, had some degree of residual contribution value, however the predictive value of the model had reached the optimal mix of variables by excluding these.

V. Optimizing the Strategy with Decision Tree Pro

The current strategy

As stated above, the current strategy for the scenario's credit grantor is that the applicant's loan-to-value be less than 100, their debt ratio must be less than 50, their FICO Score must be at least 600, and finally their in house custom score must be 200 or greater. Figure 15 shows the current decision tree.

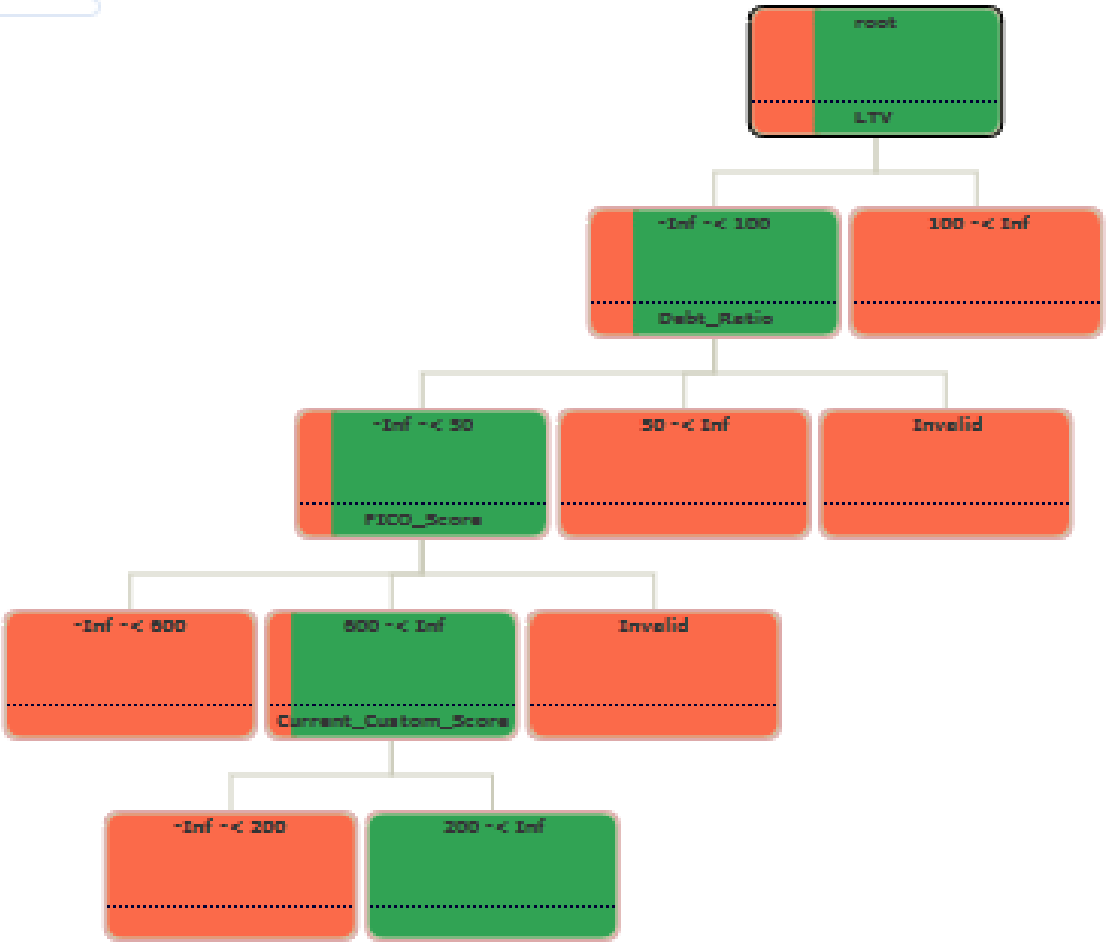


Figure 15. Current Decision Tree Strategy

This model is quite effective at identifying those who will pay their loans. Indeed, using the validation data, the current model only denies loans to those who would otherwise have paid off their loan as negotiated at a rate of 24%. At the same time, the current model collects 75% of the weighted total of possible revenues. What the current model fails to identify efficiently are those borrowers who are going to default on their loan. Again, using the validation data, the current model accepts 43% of those applicants who will default on their loan. Table 1 further shows the predictive accuracy of the current strategy. Thus, the goal for any new strategy is to reduce the ratio of accepted loan applicants who will default on their loans, while at the same time not rejecting those who would have otherwise paid as negotiated, and avoiding the loss of that potential source of revenue.

Table 1. Predictive Accuracy Measures for the Current HELOC Application Approval Strategy

Risk of Default		Interest Revenue	
LR^2	0.106	R^2	0.030
AUC	0.689	RMSE	596.121

Like the iterative process in using Scorecard Professional to determine a new scorecard, so too is the process iterative in determining the new strategy for loan approval in Decision Tree Professional. Table 2 presents model statistics of the team's different strategies using the training data.

Table 2. Model Statistics for Decision Tree Outcomes with Chosen Model Highlighted Yellow

Tree	Risk of Default		Interest Revenue	
Name	LR^2	AUC	R^2	RMSE
Model 1	0.170	0.735	0.063	596.407
Model 2	0.096	0.670	0.046	601.800
Model 3	0.178	0.755	0.064	595.999
Model 4	0.139	0.727	0.054	599.310
Model 5	0.232	0.806	0.080	590.809
Model 6	0.196	0.777	0.379	485.675
Model 7	0.131	0.716	0.354	495.124
Model 8	0.207	0.791	0.384	483.553
Model 9	0.209	0.786	0.438	461.924
Model 10	0.216	0.791	0.067	594.946
Model 11	0.198	0.773	0.102	583.711

As can be seen from Table 2, focusing on improving the predictive accuracy of identifying those who would default is typically at the expense of the predictive accuracy of revenues. This is due to the fact that restricting the acceptance levels for the multiple variables in the strategy in order to better identify those who will default will typically also reject those who would have paid as negotiated, resulting in the loss of their collected revenue. Conversely, as can be seen from the current model, focusing on accepting those who will pay typically results in also accepting an unfavorable rate of those who will default. Thus these models take in a great deal of revenue, but they also carry an unacceptable level of risk which could ultimately lead to lost revenue and also the invested loan principal.

VI. Results and Recommendations

Model Ten from Table 2 was chosen as the final new strategy to be employed by the credit grantor. The new strategy is as follows:

- If the loan-to-value is 90 or greater, then the applicant would have been dropped in the other models. In the new model, an applicant's new risk score is assessed in order to reduce the rate of rejecting those who would pay as negotiated. If the applicant has a new risk score of 4.517 or greater, the applicant is approved for a medium priced loan. If the applicant's new risk score is between 3.720 and 4.517, then the applicant is approved for a high priced loan. If the applicant's new risk score is less than 3.720, then the applicant is rejected.
- Additional variables are still assessed in order to identify applicants who will be approved for a low priced loans. These steps were taken to better refine the model, otherwise the credit grantor could simply use the new risk score developed by the team. These additional variables further allows the credit grantor to identify those applicants who are eligible for low, medium, or high priced loans.
- If the loan-to-value is less than 90, then the applicant's debt ratio is also assessed.
- If the applicant's debt ratio is 40 or greater than the applicant must have a new risk score of 3.810 or greater to be approved for a high price loan. If the applicant has a new risk score of less than 3.810 then the applicant is rejected.
- If the applicant's debt ratio is less than 40 then the applicant's FICO Score is assessed.
- At this point, the applicant must have a FICO score of 650 or higher to continue in the application approval process. If the applicant has a score of less than 650 then the applicant is rejected.
- If the applicant has a FICO Score of 650 or higher, then the applicant's new risk score is assessed.
- If the applicant has a new risk score of 5.012 or higher than the applicant is approved for a low price loan. If the applicant's new risk score is between 4.234 and 5.012 then the applicant will be approved for a medium priced loan. If the applicant has a new risk score of less than 4.234 then the applicant is rejected.

Figure 16 shows the new strategy.

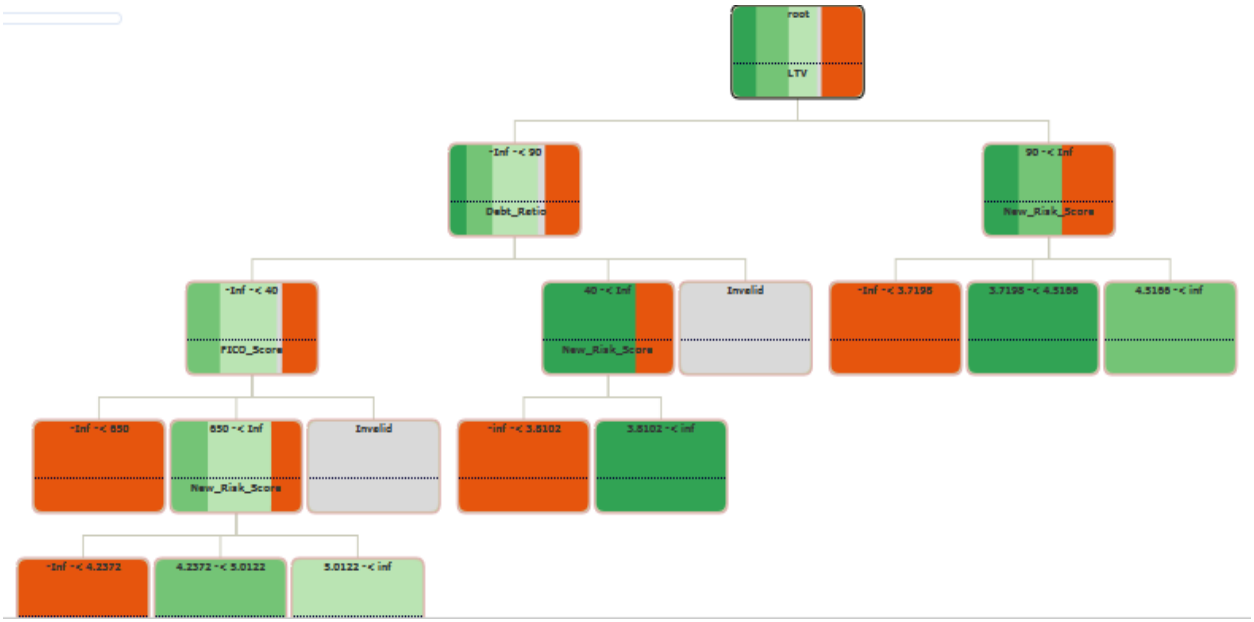


Figure 16. The New Decision Tree Strategy

Restricting the levels of the loan-to-value, debt ratio, and FICO Score relative to the current model, allowed the new model to correctly reject those applicants who will default. Indeed, using the validation data, the new model rejected 77% of those applicants who would have defaulted. Adding the additional steps between loan-to-value and debt ratio and between debt ratio and FICO Score allowed for the model to not reject an unacceptable percentage of those who would have paid as negotiated. Using the validation data, the new model only rejected 30% of those who would have paid as negotiated and collected 66% of the weighted total of all possible revenues. Thus, the new model is substantially better at predicting those who will default on their loans relative to the current model, while at the same time not losing substantial levels of revenue because of erroneously rejecting those applicants who would have otherwise made their payments as negotiated. This new strategy, using the new risk score developed in Scorecard Professional, should be utilized as it mitigates the risk of lost revenue and principal, due to accepting applicants who will default on their loans, while not rejecting an unacceptably high quantity of those applicants who would otherwise have made their payments in a timely manner.

VII. Conclusion

Our Final Scorecard Model with an LR^2 of 0.241 and ROC of 0.817 along with a perfect contribution distribution across the step-one and step-two variables proved superior to the Current Complete Custom Score model with an LR^2 of 0.203, and ROC of 0.791, which were the benchmarks to beat. Our final 10 variable model is better able to predict the outcomes of the dependent variables 'paid-as-negotiated' and 'interest revenue collected to date' in choosing which borrowers to approve and which to decline than the Current Complete Custom Score model. Furthermore, our Decision Tree was able to capitalize well on those borrowers who would continue to pay as negotiated, by using our strategy to offer loans to our approved customers with a tiered rate structure, based upon the risk profiles created by using our final model.

The Current Complete Custom score model is inferior to our Final Scorecard model since it used both the Current_Custom_Score and the FICO_Score as redundant highly correlated variables, which caused it to suffer from multicollinearity, which causes large standard errors in the related independent variable coefficients, leading to type II errors and over fitting of the model. The FICO and Custom Score model created by the lender we were assisting likely use some of the same or at least similar variables obtained from a borrower's credit bureau. ("Multicollinearity.").

Introducing highly correlated variables into a model will cause the coefficients to swing negative to correct for the magnified duplicative effect of using the same variables more than once. The model thus loses its predictive power as it becomes highly sensitive to minor changes due to the increasing variance of the coefficients to account for this amplification caused by the highly correlated variables. Removing the duplicated variables would correct this shortcoming, however in the Current Complete Custom score, the correlated variables are buried within the variables FICO_Score and Current_Custom_Score, where they are not easily separated out due to the proprietary nature of each scoring model.

Our Final Scorecard model was able to beat the Current Complete Custom score primarily due to the fact that we carefully evaluated each variable as we added them into the model, and we do not appear to suffer from multicollinearity since our variables are each unique and are not highly correlated to each other. Each variable we included in our final model increased the functionality of the model as a whole, thus, with the proof of the higher LR^2 score of 0.241 and ROC score of 0.817, our final model demonstrates that we succeeded as it is clearly superior to the original model we were asked to evaluate and improve upon.

VIII. References

"Credit Card Glossary: Terms and Definitions." *CreditCards.com.* http://www.creditcards.com/glossary/.

"Credit History." *Wikipedia: The Free Encyclopedia* http://en.wikipedia.org/wiki/Credit_history.

Culhane, Patrick. "Method and Apparatus for Scoring the Likelihood of a Desired Performance Result." 2003: n. pag. Print. http://www.google.com/patents/US6513018.

"Debt-to-income Ratio." *Wikipedia: The Free Encyclopedia.* http://en.wikipedia.org/wiki/Debt-to-income_ratio.

Detweiler Gerri. "What is a Good Credit Score." *Credit.com.* http://www.credit.com/credit-scores/what-is-a-good-credit-score/.

"Equifax Credit Report Format." *Mer*it *Credit Engine.* http://creditengine.net/equifax_format.htm.

"Experian Credit Report Format." *Mer*it *Credit Engine* http://creditengine.net/experian_format.htm.

"Home Equity Line of Credit." *Wikipedia: The Free Encyclopedia.* http://en.wikipedia.org/wiki/Home_equity_line_of_credit.

"Key Dimensions and Processes in the U.S. Credit Reporting System: A Review of how the Nation's Largest Credit Bureaus Manage Consumer Data." *Consumer Financial Protection Bureau.* http://files.consumerfinance.gov/f/201212_cfpb_credit-reporting-white-paper.pdf.

"Multicollinearity." *Wikipedia: The Free Encyclopedia.* http://en.wikipedia.org/wiki/Multicollinearity#Consequences_of_multicollinearity.

"Second Mortgage." *Wikipedia: The Free Encyclopedia* http://en.wikipedia.org/wiki/Second_mortgage.

Simon, Jeremy M. "FICO's 5 Factors: The Components of a FICO Credit Score." *CreditCards.com*. http://www.creditcards.com/credit-card-news/help/5-parts-components-fico-credit-score-6000.php

"TransUnion Credit Report Format." *Mer*it *Credit Engine.* http://creditengine.net/transunion_format.htm.