Avoiding a Knock-out

The Danger of Models Built Out of Context (And How to Recover)

ensemblex

INTRODUCTION: A ROLLOUT GONE BAD

Client is a consumer lender operating across several markets in Latin American with more than $200 million in assets. The company's core product is personal loans, with loan sizes ranging from $400-$15,000.

Prior to 2018, Client used a generic "off-the-shelf" credit score as their primary tool for underwriting. In early 2018, they engaged a well-known data and modeling company to build a custom origination score for them. The company shared analysis and validation of the new custom origination score, demonstrating substantially better performance than the incumbent model. The new model offered the potential to increase business significantly without increasing risk. Client launched this promising new custom origination model in mid-2018.

A month after launching the new model, Client could already see something was seriously wrong. Early payment delinquencies spiked. Client was forced to quickly scale back originations and apply stopgap rules-based overlays to keep their portfolio risk from spiraling out of control.

What went wrong? Prior to launching the new custom model, Client had applied various rules-based decisions, called knock-outs ("KOs"), as part of their underwriting process. These KOs came and went over time, leaving an inconsistent dataset available for model development that wasn't representative of Client's latest applicant base. The new model delivered to Client didn't account for this vital background, and the accompanying implementation plan lacked clarity on how these KOs should continue (or not). Client's new model looked great within the narrow confines of the development data, but it was blind to the limitations of that data and how it applied to Client's current business.

MORE KOS STOP THE BLEEDING

Fortunately, Client had a vigilant analytics team, who spotted very quickly that something had gone seriously wrong with their new model rollout. To stem the damage, the Client analytics team used univariate and multivariate analysis to spot pockets of risky borrowers, introducing even more KOs to their underwriting process to deny similar applicants in the future.

The Client team was able to meaningfully rein in risk, but at a cost. Rules-based cuts like KOs are necessarily crude, broad tools that eliminate entire groups of applicants. This not only leads to sub-optimal risk differentiation, but it also creates a more clunky underwriting process and constrains business growth.

Client turned to Ensemblex to build a better underwriting solution. Our mandate was to build a more powerful machine learning ("ML") underwriting model which could

further control Client's losses, simplify their underwriting process and grow their business. To do all of this safely and effectively, we grounded our model development, analysis, and rollout strategy in Client's business policies and context.

NAVIGATING KO ROLLBACK

Our aim was to find the balance between simplifying the underwriting process by rolling back some KOs, while avoiding blind spot risk, the very issue that blew up the previous model rollout. There is no single, universal tactic or trick which alone solved this puzzle for us. Instead, our approach was honed by years of experience, a collection of analytically rigorous methods, careful examination of available data and consultation with the Client team.

Prioritize and lean on our client's expertise

At the time Ensemblex began our model development for Client, there were 19 different KOs in the full underwriting funnel. These ranged from standard demographic cuts (e.g., an age minimum) to basic credit risk cuts (e.g., not currently more than 60 days delinquent on another account at time of application) to very specific "segment" KOs tied to the incumbent model (e.g., not less than 30 years old unless meeting specific bank and score criteria).

Rolling back all of these KOs at once would have been very risky. Many are intuitive predictors of credit risk that had been in place for years, meaning that the model development population contained no performance data from applicants who didn't pass these KOs.

Instead, we consulted with the Client team to prioritize which KOs were most important to them for initial rollback. Together, we chose seven to eliminate in conjunction with our model's implementation. These choices were based on the Client team's knowledge and experience, as well as the KOs' dependence on incumbent model scores and the ability to use natural tests in the data to explore blind spot risk (see "Using natural tests in the data" below).

Analyze and structure the available data

Analyzing and structuring available data were critical to both better understanding blind spot risk and creating the appropriate evaluation data.

Blind spot risk

We examined the historical application of KOs to understand which presented true blind spot risk. KOs that had been applied inconsistently over time often had some loans with outcomes in the development dataset and thus presented less blind spot risk, making them more appealing candidates for rollback. On the other hand, KOs which

had been in effect for the entirety of the model development period represented true blind spots, as the model had no exposure to the impacted populations. While we didn't completely disregard such KOs from rollback consideration, we approached them much more carefully.

Appropriate evaluation data

When analyzing model performance and impact, we worked closely with the Client team to replicate all current KOs in the dataset as thoroughly as possible. Then, once KOs we prioritized for rollback versus continued use, we ensured that we matched this logic in our evaluation dataset. Applicants who would continue to be KO'ed were excluded from our analysis of our model's performance. We instead focused our swapset and economic impact analysis on those applicants who would actually be eligible for approval by our model – i.e., passing all new selection criteria.

This step of replicating credit rules in your evaluation dataset may seem quite simple and intuitive, but we have found that it is too often neglected or done haphazardly when analyzing and validating model performance.

Use natural tests in the data

The introduction and/or removal of KOs created an opportunity to use empirical performance to examine the risk of rollbacks. We call these opportunities "natural tests" in that they weren't designed as performance tests *per se,* but rather were indirectly created by inconsistent application of rules over time.

In particular, we were interested in KOs that had been introduced more recently, but were not in effect for the full duration of our dataset. In these cases, we could isolate historical loans with performance data that would no longer have been approved under Client's current framework. We then examined these loans to validate that our model could still slope risk in these populations and that this risk-sloping was consistent with the non-KO population. This allowed us to confirm that our model could work in both populations, increasing comfort with the proposed changes.

Scrutinize swapsets

As part of any model implementation, we conduct a thorough swapset analysis – a careful study of where our new model differs from the incumbent model and associated rules and what is driving those differences.[1] In the case of Client's model, several

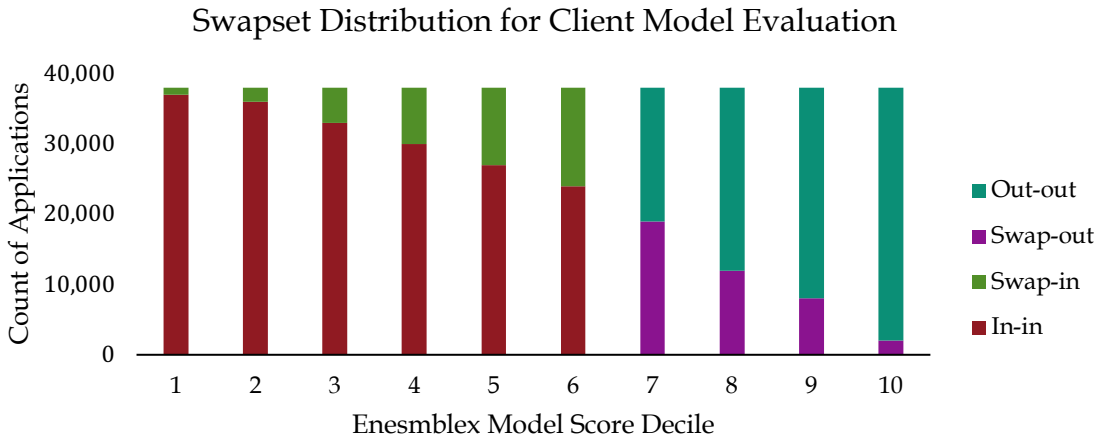
[1] The basic segmentation used for swapset analysis consists of: "in-ins," applications approved by both the old and new models; "swap-ins," applications approved by the new model but denied by the old; "swap-outs," applications approved by the old model but denied by the new; and "out-outs," applications denied by both models.

fundamental aspects of swapset analysis allowed us to further build confidence in our plan.

Sizes, distributions and top features

Our swapset analysis starts by understanding just how different our model's decisions are from incumbent decisions. We do so by examining swapset sizes, distributions and top feature values. While we expect a healthy swap-in population relative to in-ins, we want to ensure that the size of swap-in population is reasonable given how the business is performing today. Client had recently done a good job reining in risk, so we expected our model to agree, more often than not, with incumbent decisions. And we expected instances where both models agreed to skew toward our best scoring applicants, while cases where we disagreed skewed toward more marginal approvals.

In this case, both the Ensemblex and Client teams were comfortable with the size of swapsets (our model agreed with incumbent approvals 84% of the time) and distributions (swap-ins overwhelmingly skewed toward the riskier scores).



Additionally, we examine the values for top model features across all swapsets. Our top model features are intuitive credit predictors that help get us and our clients comfortable with what is driving model performance. When examining these features across swapsets, we expect the swap-in population to have better values than the swap-outs, but not as good as the in-ins. Examining top features for our model, we found this to be the case.

Swapset Values for Top Model Features

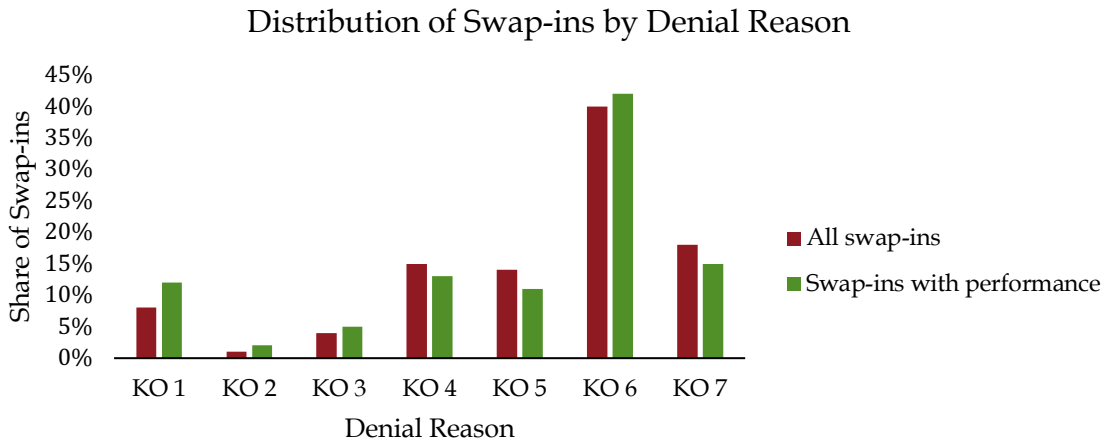
Variable	In-in	Swap-in	Swap-out	Out-out
Positive references on file (mean)	1.2	0.9	0.3	0.3
Max days DQ (percent > 0)	6%	11%	14%	22%

Application channel (pct outbound)	27%	17%	15%	7%

Using empirical performance

In the case of Client's model, we had enough empirical swap-in data from the natural tests discussed above that we could confidently project swap-in performance even in the areas where KOs were rolled back.

To ensure that the empirical performance we had available was representative of all swap-ins, we compared the distribution across KOs of swap-ins with performance to the full swap-in population. We concluded that the instances of performance we had for swap-ins were representative of the full swap-in distribution across the KOs, building further confidence in our ability to roll them back.



Deep-dives into concerning trends

Finally, we explored values and distributions for other key segments in the swapsets for any potentially troubling trends. Two we spotted were that our model was swapping in both more non-bank-upgrade[2] and younger borrowers. Given that we expected non-upgrade and younger borrowers to be riskier (relative to upgrade and older borrowers, respectively), this was problematic.

Examining these segments more closely, however, we found that *overall* our model still agreed with the incumbent model in the overwhelming number of cases, finding upgrade and older borrowers less risky. However, it was able to isolate toxic subsets

[2] Bank upgrade is a custom identifier created by the Client analytics team using several key attributes within the banking data. In general, customers flagged as "upgrade" are expected to have better risk profiles.

within the upgrade and older populations, and replace them with worthwhile non-upgrade and younger borrowers, as the table below demonstrates.

Swapset Performance for Age and Bank Upgrade Segments

Variable	Segment	Bad rate
Age	Swap-in < 30	12%
	Swap-out >= 30	23%
Bank classification	Swap-in non-upgrade	13%
	Swap-out upgrade	21%

THE RESULT: A ROLLOUT GONE RIGHT

Final analysis of our model's performance, conducted jointly with the Client team, showed its potential to reduce risk by 11% relative to the current rules-based waterfall and 40% relative to the period immediately after the previous model rollout. We achieved these gains while simplifying their underwriting process by eliminating 7 of 19 KOs immediately upon rollout of our model.

Six months after rollout, the model's risk indicators are even better than projected. Client's business is well on track to ramp back up to pre-covid volumes more quickly than expected, and they are looking at further opening up approval rates. They are also well-positioned to begin testing cycles to rollback additional KOs.

Client's previous model rollout left them with drastically increased and unexpected risk levels and the need to apply additional KOs, which further complicated their underwriting process and reduced their eligible applicant universe. By taking the time to fully understand the business context and policies, Ensemblex ensured that this latest rollout told a different story. Client now has a business with improved risk and poised for further growth. And, in our model score, they have a lever they can trust to drive that growth.

THE ENSEMBLEX DIFFERENCE

At Ensemblex, we are uniquely positioned to help businesses unlock the power of ML. Our elite data scientists work with clients to apply the latest in ML techniques and technologies, but we anchor our work with a deep and collaborative understanding of our client's business. No matter how powerful the algorithms, a model that is unmoored from the underlying business and its data is at best sub-optimal and at worst dangerous.

Ensemblex is a specialty analytics partner for neobanks and fintechs. We are a team of strategic, data-driven lending experts who have held executive roles at companies like

Capital One, CitiBank and ZestFinance. We have decades of experience working with companies ranging from leading banks, smaller lenders and fintechs across multiple asset classes including personal loans, credit card, auto, SMB and student. Our team was among the first to actively use ML techniques for consumer underwriting, and we have been building and implementing ML models at our own companies and for our clients for over a decade.

Learn more and meet our team at www.ensemblex.com or contact us at info@ensemblex.com.