vestigoanalytics.com
info@vestigoanalytics.com

Machine Learning and Credit Risk: Beyond the Buzz

Machine learning and artificial intelligence are undoubtedly among the hottest topics in credit risk today. We discuss their usage in risk modelling and attempt to separate the reality from the marketing.

As machine learning and artificial intelligence have become increasingly prevalent in finance, lenders look to their budding usage in decisioning and default prediction. Applying this technology to credit risk modelling has huge potential but also provides unique challenges which must be overcome to ensure success. However, despite much literature on the issue, articles and papers often fail to take into account the practical considerations of building and implementing such models in a credit risk setting.

What is it?

Before too much discussion, we should clarify what we mean by machine learning and artificial intelligence. Often these terms are used interchangeably, although they are distinct concepts.

Despite conjuring up images of machines with super-human intelligence, artificial intelligence (AI) is simply a set of techniques and rules which allows a computer to mimic human behaviours. Everyday examples of AI in action are email spam filters, call routing software and customer service chatbots. Machine learning (ML) is a subset of AI where algorithms use data to learn to do a pre-specified task, without having to be told how to do it.

Our focus within this paper is the application of machine learning to credit risk, and in particular the creation of default prediction models. Fraud detection, transaction approval and pricing are other common areas of ML application, but default prediction has the advantage of being easily understood and allows direct comparison to traditional credit risk scorecards. Numerous algorithms have been tested, including regression methods, random forests, gradient and adaptive boosting and deep-layer neural networks. With a huge array of algorithms available an exhaustive assessment is difficult, but the breadth of tools considered should support the generality of any findings.

Doing the Groundwork

In any predictive modelling, often the most time-consuming step is not building the model itself but preparing and understanding the data that goes into it. Variables must be cleaned and processed, with any data quality issues identified and remedied. Some advocates claim machine learning offers an appealing short cut: algorithms will pick up underlying trends on their own, without the need for careful data handling or extensive preparation. However, while it is possible for most popular algorithms to run with minimal data preparation (usually just missing value imputation and scale standardisation), significant performance uplifts can usually be achieved with a more comprehensive approach.

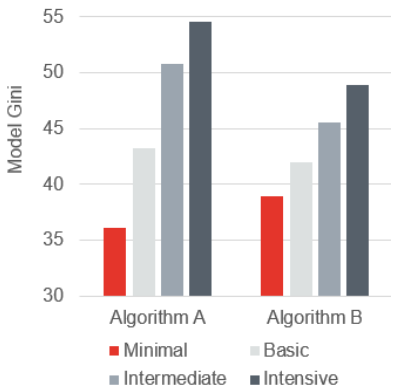

Figure 1. Impact of different levels of data preparation on average model Gini for two algorithms. The benefits of better data preparation often outweigh the uplift available from switching algorithm.

Elements that need to be considered when preparing datasets for machine learning include outlier detection and handling,

Vestigo Partners Limited
Registered in England and Wales. Company No. 10826517

normalisation, data transformations and new variable creation. Many algorithms also do not interpret raw text well without significant guidance, so there are benefits to transforming this to numeric data, often through encodings or using dummy variables. An example of specific relevance to credit risk modelling is credit bureau data, which may have many different categories of 'missing' data and numeric default values that should not be interpreted ordinally.

The best approach to data preparation and the level of benefit derived will vary on a case-by-case basis. However, when comparing different levels of pre-processing, from bare minimum to intensive, our analysis shows that the improvements here can overshadow differences between algorithms, particularly at the top end of performance.

Some algorithm classes, such as random forest classifiers, are more resilient to data quality issues, although some impact remains. Random forests are large collections of randomly generated decision trees that make an overall prediction using a weighted average of their individual scores. As decision trees can split on any variable value, they (and random forests as a result) are relatively insensitive to the type and scale of the data. Consequently, random forests can be used prior to any intensive data preparation work to quickly identify predictive variables – the more frequently a variable is used in the forest, the more predictive it is. This allows attention to be focussed on the very best variables to ensure they are utilised effectively.

Garbage In, Garbage Out?

Machine learning is often said to require no outside knowledge, as the algorithms teach themselves. Whilst this is largely true for the actual algorithms (how many modellers really need to know the finer details of Tikhonov regularisation of support vector machines?), it is still necessary to understand the data being used and the benefits and limitations of the models being produced. Without understanding how data has been sourced and the operational processes surrounding it, it is very easy to make mistakes and feed the models data that they cannot be expected to interpret correctly.

As a very simple illustration of this, let's consider an example of sample bias. Datasets used for default prediction are often based on empirical data resulting from previous lending activity. As such these are impacted by previous operational and selection biases such as cherry-picking (for the sake of this paper we will avoid the interesting topic of reject inference!).

In this case, the lender in question did accept customers with historic bankruptcies, but with additional restrictions and only after careful review by an experienced underwriter. As a result of this, the performance of bankrupt cases was actually relatively strong, with bankruptcy appearing as a positive attribute on the development dataset. Lacking this contextual awareness, most of the ML algorithms tested ranked new bankrupt cases as higher quality than the general population.

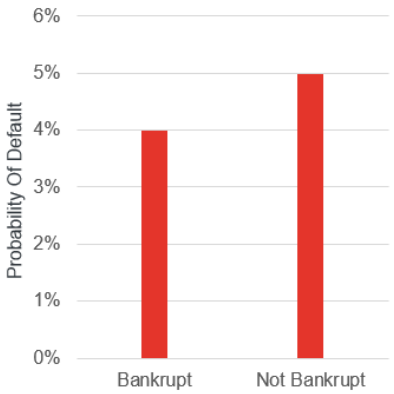

Figure 2. Average predicted probability of default across all algorithms tested for bankrupt and non-bankrupt populations. The bankrupt population was predicted to be 20% less likely to default.

If not spotted and corrected for prior to model deployment this could potentially be a costly issue, with the performance of new applications including bankruptcy being over-estimated. Whilst experienced machine learning practitioners with good credit risk knowledge will avoid pitfalls such as this, it is easy to trip up if the correct level of care is not taken.

Vestigo Partners Limited
Registered in England and Wales. Company No. 10826517

Show Your Workings

In the previous bankruptcy example, this model was actually implemented by the lender, but the issue was spotted early by underwriters involved in case reviews. As laid out previously it seems indefensible that this was missed, but the ML model used by the business's data scientists was not easy to interpret.

Model transparency is particularly important in credit scoring as decisions produced by the model need to be explainable and justifiable, to customers, internal stakeholders and regulators. The development data may (coincidentally) show that Bills are less risky than Bens, but this isn't a trend likely to hold in the general population, even if the resultant model has a marginally higher testing Gini because of it.

Such 'intuitive' insights into what is and isn't suitable are easy for human data analysts. However, ensuring this type of behaviour in an algorithm is much more difficult. Typically, machine learning algorithms are data hungry – they will use all the variables provided in some way, unless they are specifically penalised for doing so. The result is that both the data going into the model and the outcomes being produced need to be analysed to ensure that decisions are reasonable.

In a similar vein, we have encountered machine learning credit risk models that have indirectly used gender, simply because it was derivable from the "pot" of unstructured data poured into the model algorithm. With the burden on the lender to prove their compliance to regulators, it is important to keep in mind the potential issues of using more data, often in a less structured form, in algorithms with limited transparency. Luckily this is an area of significant focus in the industry and there have been some excellent steps forward in harnessing machine learning power while retaining tractability of the resulting models.

Too Much Information?

As computing power increases, machine learning allows the analysis of more data in more ways than ever before. Open banking, multi-channel interactions and the rise of social media have the potential to give institutions access to a level of detail never previously possible, with thousands of customer touchpoints to analyse.

Whilst some of this new data will undoubtedly be highly predictive, some will also be almost completely uncorrelated to risk. The adage has long been that more data is better, but in the world of machine learning this does not always hold. As well as increasing development times, including large amounts of unpredictive or weakly predictive variables can significantly reduce performance in some algorithms due to overfitting.

To illustrate this, we added variables consisting of normally distributed random data to a modelling sample, which was then split into train and test data as usual. A separate out-of-universe sample was retained for the purpose of measuring the 'real world' performance of the developed models. Almost all algorithms tested saw reductions in average Gini produced with the addition of random data, although some were significantly more resilient than others. Overall, we found that an increase of 50 uncorrelated variables reduced the average Gini produced by 2.3%.

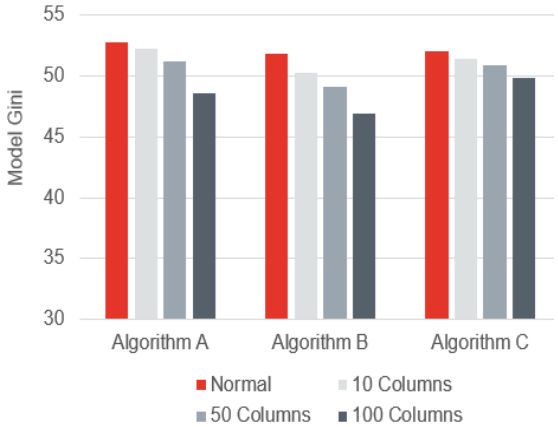

Figure 3. Impact on performance Gini of adding uncorrelated variables to development sample for several popular algorithms.

Using hundreds or even thousands of variables in a model can also cause issues with model stability monitoring after deployment. Any significant change in the distribution of a variable, whether resulting from population

Vestigo Partners Limited
Registered in England and Wales. Company No. 10826517

drift or changes at the data source, may have a negative effect on model performance. As more variables are used, the likelihood of a significant distribution shift in at least one of those variables rises. The impact of such changes is difficult to monitor as complex variable interactions can create chaotic and unpredictable results.

Our analysis clearly shows then that more variables aren't always better for model performance if those variables are largely noise, and especially if the model is not carefully supervised. This is a very interesting finding and flies somewhat in the face of the usual 'big data', more-is-better mantra common in the data analytics world. Extra sources of predictive data will clearly be beneficial, but there are risks to simply indiscriminately throwing more data at an ML modelling algorithm. Fortunately, there are many well-tested variable reduction methods and removing undesirable variables during data preparation, prior to development, will alleviate these issues.

Whereas additional variables should be treated with caution, more independent data points almost always have a positive impact. Using random forest classifiers, we developed risk models on development samples of various sizes, from 100 through to 100,000 loan records. Machine learning datasets can run into millions of records, but samples larger than 100,000 cases become less common for lenders when factoring in recency of data and costs such as retrospective credit bureau data. Increasing sample size led to both increased predictive power and a reduction in the standard deviation of model performance; the average model Gini almost doubled between 1,000 and 100,000 record samples.

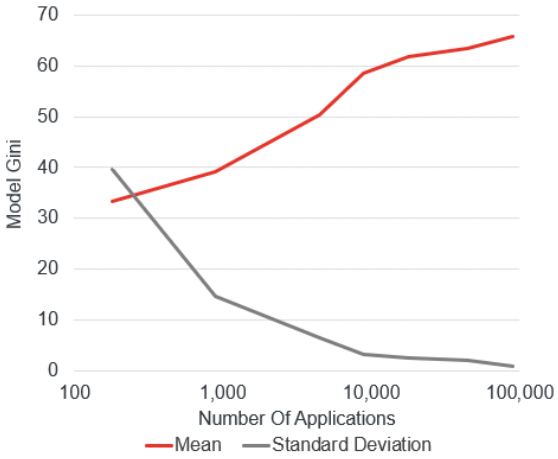

Figure 4. Average gini and standard deviation for models produced by random forest classifier with varying numbers of applications in development sample.

Shaping the Future

Machine learning is a powerful technology and those businesses that temper data science techniques with pragmatism and credit risk know-how will reap significant benefits. Areas such as pricing, limit setting, fraud detection and transaction authorisation are ripe for the use of such models. Standard credit underwriting modelling requires a little more care, due primarily to regulatory and transparency considerations, but we expect to see growing prevelance and benefits from machine learning models applied in this space.

The biggest risk in the successful implementation of these techniques is the 'silver bullet myth'. Machine learning is complimentary to, but does not replace, core competencies such as understanding your data, your operation and your market. Paradoxically, those lenders most likely to come unstuck with machine learning are those with the highest expectations of it; those who think it can replace knowledgeable people, good data and considered modelling. For the rest of us, this is a fantastic toolset that will let us do our jobs more effectively, serve our customers better and manage our businesses with improved clarity.

Vestigo Partners Limited
Registered in England and Wales. Company No. 10826517

Written By:

Natasha Conradi
Consultant

About Vestigo

Vestigo is a team of experienced and dynamic analytics and credit risk professionals established in 2017. With a strong pedigree in financial services and credit risk, and experience across numerous other industries, we deliver comprehensive analytical support and risk services. Our London-based team of consultants, specialists and practitioners provide services to clients worldwide on a consultancy, contracting or outsourced basis.

To find out more, please contact Paul at paul.matthews@vestigoanalytics.com or call 07391561015.

Vestigo Partners Limited
Registered in England and Wales. Company No. 10826517