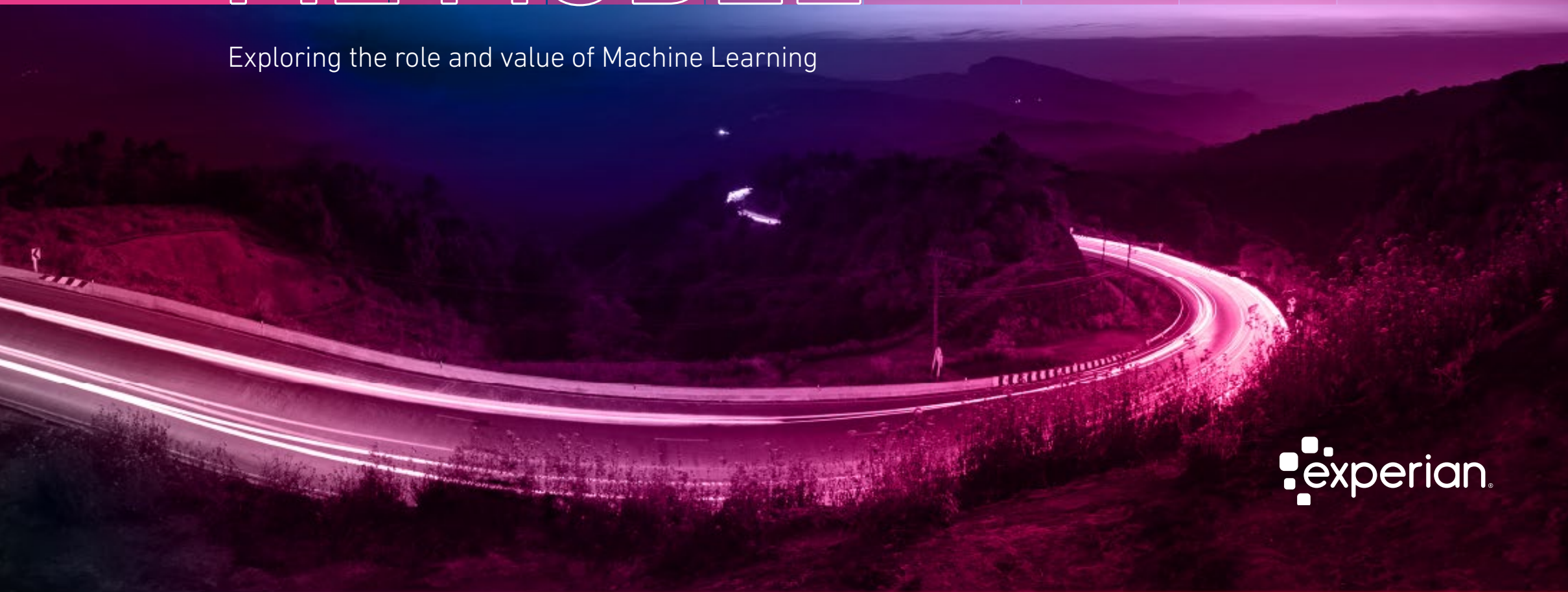
Journey of an ML MODEL

Exploring the role and value of Machine Learning



experian

Lack of clarity over the role and value of Machine Learning

With so much media hype about Artificial Intelligence (AI) and Machine Learning (ML), it can be difficult to cut through the noise and understand the true value of this technology.

It is only through ML that businesses can distil the ever-increasing quantities of alternative and unstructured data to reveal powerful predictive insights. Many businesses are already using ML as a means of creating accelerated and positively differentiated data-to-insight-to-action pathways across multiple use cases.

This guide explains our approach to ML by taking you through eight key stages in the process of ML model development and usage. As we progress through the journey of an ML model, we focus on the role and importance of the enabling capabilities and technology in helping maximise and scale the widest possible range of positive value outcomes.

Experian has over two decades of experience working with ML and has successfully developed and implemented thousands of ML models. We understand the potential risks associated with this technology and have an unwavering focus on safely creating value from ML.

LEVERAGING
SPEED AND SCALE

ENSURING
ACCURACY, CONSISTENCY, AND AGILITY

ACCELERATING
INSIGHT CREATION AND APPLICATION

Define the business problem

The first step in leveraging the value of ML is to understand the specific business problems that need to be solved.

These can be extremely wide-ranging and are often prioritised based on the nature of the required business benefits. In many instances, business problems are closely related and this creates further challenges in terms of identifying the optimal balance between potentially competing priorities.

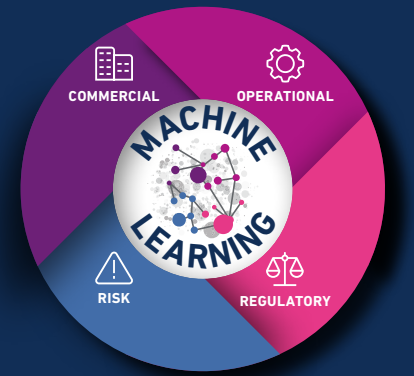
> "A failure to manage business problems holistically can result in ML models that lack sufficient scalability, efficiency and overall risk agility.
>
> To avoid this, close alignment between different insight creation requirements is needed across multiple business functions."

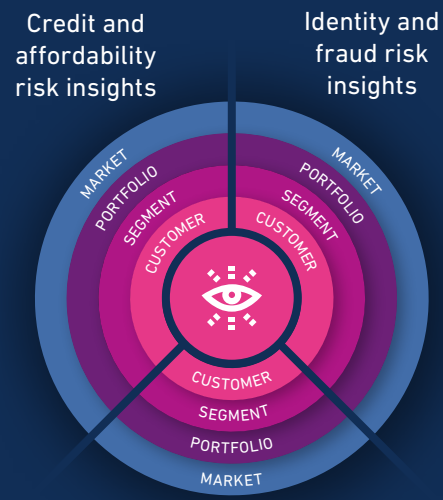
Identifying the business problem helps to determine where ML can be applied to improve predictive accuracy and gain greater insight.

The main areas are:

- **Credit and affordability insights**
  Gaining a complete view of customers' overall level of indebtedness, usage of credit and their ongoing ability to pay.
- **Identity and fraud insights**
  Helping establish whether the customer is a real person who owns the identity that is being presented.
- **Marketing and value insights**
  Assessing the overall potential value of the customer and their propensity to buy, churn or indulge in potential financial risks such as excessive gambling.

Some insights can be applied at either a customer, segment, portfolio or market level - helping ensure a complete view of the changing nature of risk and value.





We have over 25 years of experience working with clients on ML projects – covering a range of business objectives

COMMERCIAL

- ✓ Understanding customer behaviour to enable more accurate marketing and support
- ✓ Accelerating customer onboarding
- ✓ Widening customer eligibility

RISK

- ✓ Earlier anticipation of emerging risks and threats
- ✓ Increasing risk understanding for new thin file applicants
- ✓ Improving credit, affordability, identity and fraud risk predictiveness
- ✓ Assessing historical declines to improve risk policy

OPERATIONAL

- ✓ Automating risk decisioning processes
- ✓ Creating early warnings of vulnerable existing customers, before they move into collections
- ✓ Reducing manual reviews without increasing losses

REGULATORY

- ✓ Minimising portfolio volatility and levels of non-performing loans
- ✓ Ensuring compliance with regulators' standards of accounting, auditing and model explainability
- ✓ Ensuring speed and accuracy of regulatory reporting
- ✓ Aligning operational processes with regulatory risk reporting
- ✓ Minimising levels of financial provisioning

Search and acquire data

Internally Sourced Data

- Transactional data MORE+
- Call centre voice interactions MORE+
- Customer behaviours and decision outcomes MORE+
- Known fraud MORE+

Customer Supplied Data

Submitted in application

- Basic customer details MORE+

Actively supplied

- Biometric MORE+

Passively collected

- Device behavioural biometrics MORE+
- Device attributes MORE+

Understanding the business problem and where insight is needed helps identify the most suitable data for analysis.

Combining a variety of data sources is the key to producing accurate ML models Multiple datasets exist and they are continually growing. Using non-traditional datasets can provide a significant uplift in predictiveness accuracy and ML is essential to gain insight from these large and often unstructured datasets.

We can enable real time access to an expanded knowledge universe made up of historic bureau data, in-house customer transactional data, and a wide range of other data assets.

Externally Sourced Data

Customer specific

- Credit bureau MORE+
- Trended data MORE+
- Trigger data MORE+
- SME Business Intelligence MORE+

Non-customer specific

- Macro-economic MORE+
- Mosaic socio-geographic MORE+
- Environment, Social and Governance (ESG) MORE+
- Fraud Consortia MORE+

Customer Consented Data

- Open Banking transactional Data MORE+

Consolidate, clean and prepare data

Data preparation involves moving and consolidating multiple types of data into a single environment. There it is cleaned, organised, structured and put in consistent formats. Data preparation can account for 80% of the total time spent on an ML project.*

Consolidation and Storage

Data staging areas need to be capable of holding extremely large volumes of data at what is effectively the start of the data-to- insight-to-action journey.

Data Cleaning and Audit

Cleaning involves identifying and removing data points that are corrupt, incorrectly formatted or missing values, rows, or columns. Duplicated data points are either merged or removed.

Additional statistical tests are also applied within and between variables to identify outliers, of unusual and illogical values to ensure data integrity.

Unstructured data

More and more of the data that is potentially available in the digital age is unstructured. Examples include the transaction histories of existing customers and the recordings of interactions between customers and call centres.

Having moved existing data into the staging area it can then be further augmented with additional external datasets.

All of this consolidated data then needs to be converted into consistent structured formats which are suitable to be used within the insight creation process.

> “Accenture estimate that **80%** of modern organisations’ in-house data consists of unstructured text, video, audio, and images.”

* Source: Amazon Web Services

Data labelling and data linking

Data labelling (also known as annotation) is the process of attaching consistent identifiers to potentially important data variables in order to create a reference dataset.

Mining customer transactions for insights

Using the labelled (annotated) reference dataset, ML models are trained to apply these labels when they ingest raw data and group variables into specific categories within a suitably designed overall taxonomy.

Banks hold vast amounts of data relating to the transactions carried out by customers.

Hidden within this data are extremely powerful behavioural insights.

If these can be successfully extracted, they can be used to create predictive models that reveal the potential for customers to buy additional products, their potential to churn and their potential to default.

HOW WE CAN HELP

- ✓ Our data and analytics platforms enable vast amounts of both structured and unstructured data to be consolidated within highly scalable secure staging environments. These also provide ready access to a wide range of additional traditional and alternative data types.
- ✓ Our specialist consultants are highly experienced in helping clients prepare, clean and create links between data sets.
- ✓ We also provide real-time categorisation services for banks’ and credit card companies’ in-house transactional account data.
- ✓ These categorisation services use ML models which have been trained using locally annotated datasets and which create categorised outputs with highly granular taxonomies.
- ✓ Our Open Banking platform manages the customer consent journey, enabling organisations of any type to gain access to categorised transactional data held with other banks.

Identify and cluster data variables

Identifying correlations against the business problem

The dataset is now ready for the data scientists to conduct an exploratory analysis of the individual data variables.

This 'feature engineering' requires the identification of those individual variables or clusters of variables that are likely to have an impact on achieving the greatest predictive uplift.

Protecting against bias

To ensure that the model does not produce results that are unintentionally skewed, the data scientist will seek to identify any variables that have the potential to create bias or inequality.

The most obvious of these are variables associated with social factors such as gender and ethnicity but socio-economic factors are equally important.

The risk of bias is mitigated by limiting ('protecting') the influence that identified variables can have on the overall model outcome. Protecting variables helps minimise the risk of inaccuracy which could have multiple implications, such as:

- Financial loss to the organisation or to its customers
- Loss of competitiveness
- Risk of accusations of discrimination
- Failure to reflect a business's Environmental, Social and Governance (ESG) strategy in areas such as equality, social and financial inclusion.

The risk of bias becomes more acute where there is limited data available.

> "1 IN 3 Data scientists believe that the social impacts from bias in data and models are the biggest problem in data science/ML"
>
> Source: '2022 State of Data Science Report', Anaconda

A good example of this is assessing credit risk for customers who have a limited credit history. This means that developing a strategy to increase the acceptance of customers falling into this category will require real care and access to all potential sources of additional insight.

Assessing predictivity

The next stage is to assess the relative predictiveness of the clusters against standard measures of probability using a snapshot of the real data that is available. This process is then repeated using a series of repeated snapshots.

HOW WE CAN HELP

- ✓ As well as providing specialist analytical consultants to help develop new client models, we also provide platform solutions that enable clients to develop and perform these activities in-house.
- ✓ As already covered within the 'search and acquire data' stage, these platforms provide access to a range of both traditional and non-traditional datasets. In addition, we also provide access to the most up-to-date, open-source data science tools within a secure sandbox environment. This enables modelling and testing with real data that is anonymised – to avoid triggering unwanted compliance or regulatory implications.
- ✓ We can also enable data scientists from different business functions to simultaneously work on the same data sets. This is important because feature engineering is a highly skilled job and will typically require data scientists to have expert knowledge of function-specific business problems. This helps maximise the speed, agility and consistency of model development across business functions.

Create Machine Learning scores

Having created the model the next step is selecting the most appropriate algorithm.

Combining business expertise with science

Algorithm selection is a highly complex task which requires considerable expertise.

Data scientists have many options and tools available to help them solve the problem at hand. However, their technical knowledge is just as important as their professional judgement, creativity and intuition.

The real skill lies in the ability of the data scientist to understand and solve the business problem in a way which reflects the subtle nuances that shape the prevailing context.

Here the term context is a reference to which data attributes are being associated with the model and which are excluded. Understanding these associations requires close collaboration between data scientists and data managers to understand the precise nature of the available data.

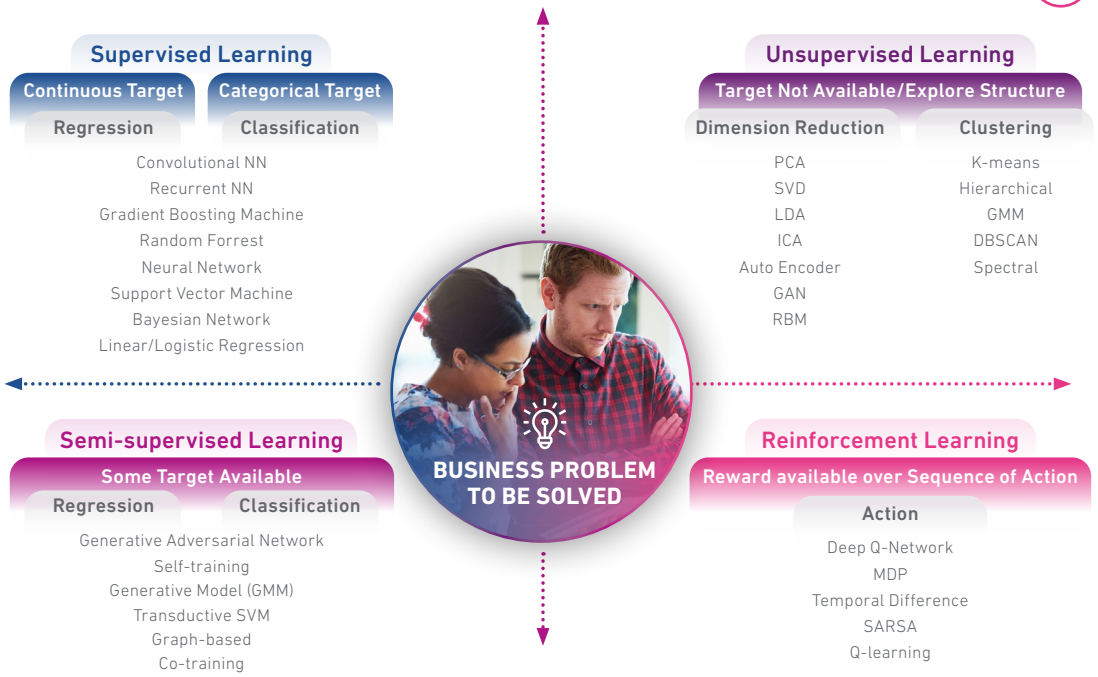
This highlights the need for clarity regarding the nature of the underlying business problem that needs to be solved.

Without a deep understanding of the context and all of its associated nuances, there is a very real risk that the business owner, risk committees, internal auditors or, at worst, supervisory regulators may not accept the model outputs.

The challenges facing organisations

Many organisations lack the ability, understanding and experience to accurately match the algorithms to the problem.

Where this is the case, organisations will be forced to rely on traditional models and scores. Without the power of ML, these organisations are likely to increasingly fall behind their competitors and suffer reduced business and portfolio quality.



HOW WE CAN HELP

- ✓ Our combination of expertise and technology means that we are able to help a wide variety of organisations with differing levels of in-house expertise and analytical capabilities.
- ✓ Organisations that lack both ML development expertise and infrastructure can access expert consultancy. We have over 100 data scientists and AI/ML specialists available to create powerful models to enhance credit, fraud risk and marketing strategies.
- ✓ We can also help more advanced organisations that are looking to develop models in-house, but who face challenges in terms of scalability and IT complexity.
- ✓ We provide data and analytics platforms to support a range of different requirements with a range of specialist tools and programmes enabling accelerated model creation and comparison.

Combine and boost Gini

Progressive organisations are leveraging the variety of data available to them by blending data variables across multiple datasets to create model performance uplift.

The power of previously unseen data correlations

The value of ML is its ability to mine vast amounts of data to uncover previously hidden insights.

Many organisations are exploring the predictive ability of non-traditional data sources. Often these new data sources contain clusters of variables with entirely fresh correlations to the business problem.

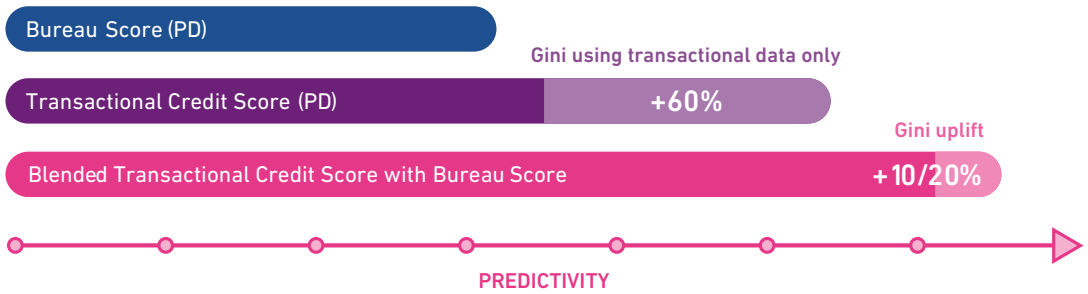
As we have already seen, these correlations can then be used to create a predictive model.

Model predictive accuracy is typically expressed in terms of the Gini co-efficient. In some instances, models using non-traditional data may have increased Gini when compared with existing scores.

An example of this is the use of transactional bank data to assess credit risk and affordability. These models typically **create a Gini uplift of around 60-70% compared with traditional scorecards.**

Blending variables to create further uplift

As the variables within new alternative datasets often have very different characteristics to the traditional data used within existing models, a further uplift can often be created by blending the variables from both the traditional and the non-traditional datasets.



Maximising decisioning accuracy across multiple datasets

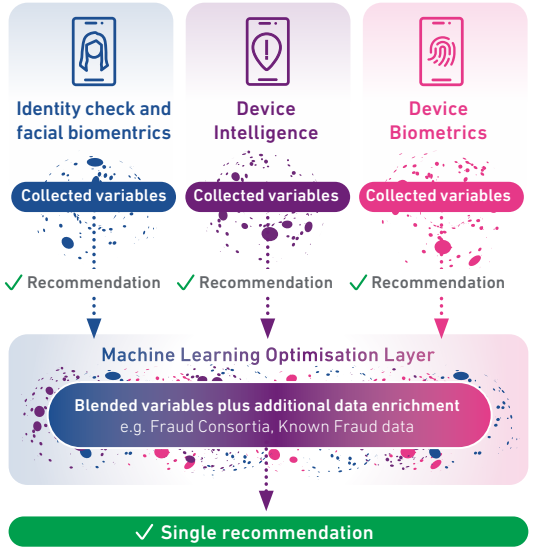
Blending data variables can add value across the customer lifecycle. In some instances, it may be a case of blending variables from two existing models. But there are also other situations where blending can also play a highly strategic role.

An interesting example of this can be seen in regard to identity and fraud. Identity and fraud risk decisioning often relies on multiple specialist assessments which form part of a multi-layered approach to fraud detection.

An ML overlay can be used to blend:

- all of the collected data variables
- the recommendations of each individual solution
- further data sets such as fraud consortia data,

and create a single optimised recommendation.

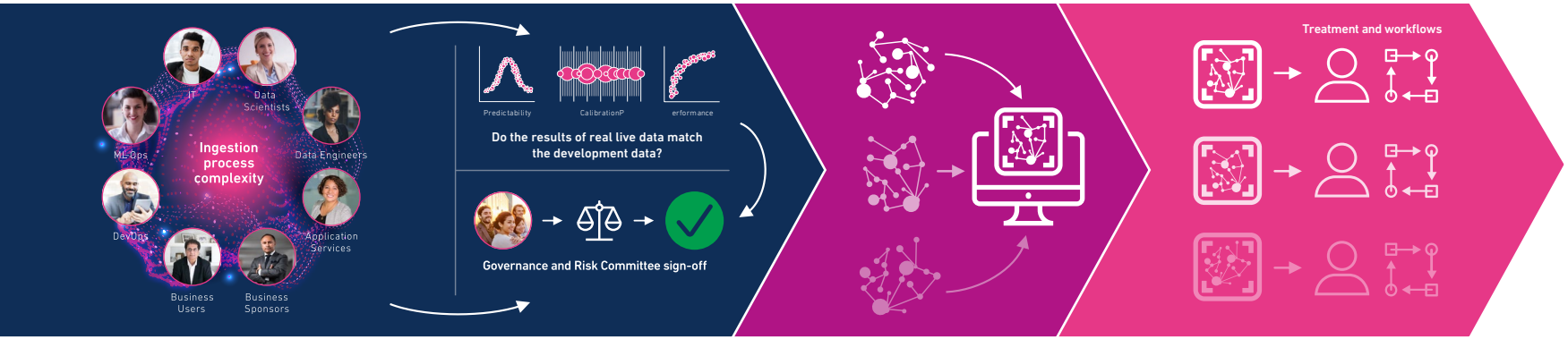


HOW WE CAN HELP

- ✓ Experian is uniquely positioned to help organisations gain an uplift in predictive accuracy across many different types of ML models for a wide range of business problems.
- ✓ We have a huge range of traditional, trended and non-traditional data sources, plus a variety of cloud-based technology solutions and the expertise to enable these datasets to be blended with existing data to get the maximum uplift in predictiveness.
- ✓ These can all be accessed from a variety of formats covering both direct and remotely facilitated delivery.
- ✓ This means that we can boost existing strategies across the lifecycle including new customer targeting and acquisition, originations risk decisioning (credit and affordability), existing customer upselling and cross-selling, identity and fraud risk management and early warnings.

Deploy anywhere

The testing and deployment of ML models represent the all-important final mile to value creation. Yet for many organisations, it is often a point of constraint.



Integrated data staging and analytics environment | **Strategy design environment** | **Production environments**

For the newly developed model to be able to create a positive return on investment, it must first be validated using real live data, signed off and then deployed into live operational processes. Unless all of these steps are completed successfully, models will remain 'on the shelf' and investments in their development will have no chance of being recovered.

Testing, Testing, Testing

Model testing can be a very complex and time-consuming process. In essence, it consists of:

- Testing the performance accuracy
- Ensuring it fully reflects the nuances and subtleties of the business problem and that it does not create any unintended consequences
- Validating performance in the live environment

Each of these considerations will often require sign-off by risk committees and this can be a time-consuming process. Model deployment will initially be into a strategy management engine, where the model can be incorporated within specific customer treatments and strategies.

At this stage the customer strategies are themselves tested – usually with the incumbent approach acting as the 'champion' and the new strategy, complete with the newly developed ML model embedded within it, being the 'challenger'.

With the strategy signed off, and the model reliably delivering accurate results it is ready to be deployed in a live production environment.

Here the models will use the continual flow of data to drive essential business applications such as applying eligibility criteria for new customers, assessing potential for fraud or assessing ongoing customer vulnerability.

The deployment challenge

The most common challenge of ML model deployment is one of integrating models developed by data scientists using Python programming language into coded IT applications within operational systems that are typically JAVA based.

The necessary work involves multiple disciplines including data, science, data engineering, dev-ops, and IT. However, incompatible skills and tools, lack of ownership and rigid organisational structures typically lead to extended timeframes adding further cost and delayed ROI.

HOW WE CAN HELP

✓ Innovation, flexibility, agility and empowerment are the consistent themes that run across our range of ML services and this is especially true of how we help overcome the deployment challenge.

✓ Our solutions enable business users to use simple point-and-click to import ML models without the need to call on the support of data science, dev-ops or IT teams.

✓ Having ingested models into the strategy design environment, decision process flows and the supporting workflow processes can be effortlessly created using drag-and-drop tools built around the model's specific use case. Here the strategies can be simulated and further tested before they go live.

Monitor, explain and optimise

Having deployed models into production, they should be subject to ongoing governance control, monitoring, validation and improvement.

Model monitoring and validation

The accuracy, reliability and contextual relevance of live models needs to be continually assessed from a variety of technical, business and compliance perspectives.

Ongoing economic uncertainty increases model risk because it increases the likelihood that models apply existing logic to new, unevaluated data scenarios. This can potentially lead to model bias, reduced decisioning accuracy, the risk of losses and potential reputational damage.

In addition, ongoing model monitoring and validation is also a regulatory requirement and the lack of appropriate resources in this area risks non-compliance.*

* "A sound validation function is crucial to ensuring the reliability of internal models and their ability to accurately compute capital requirements." (ECB 'Instructions for monitoring and validating results of internal models' - February 2019)

Explainability

Machine learning models can be notoriously difficult to explain. This can cause problems with consumers, internal auditors and regulators.

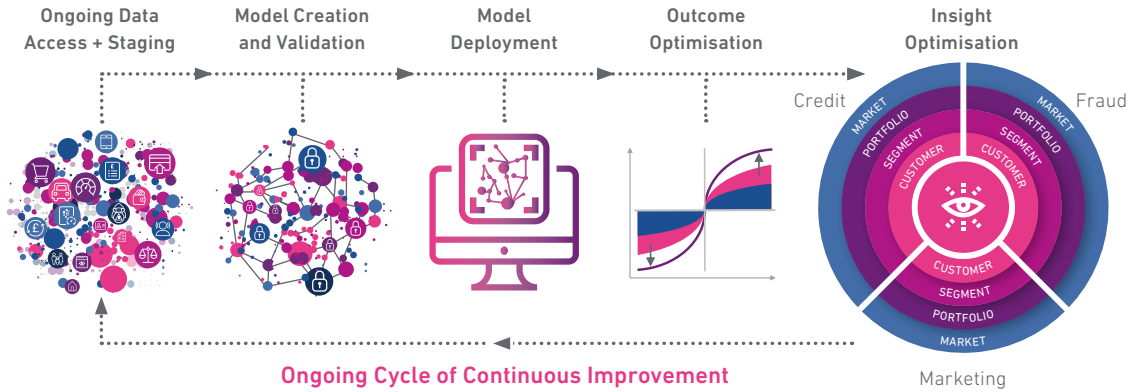
Consumers that are unable to obtain a clear explanation for an adverse decision, may feel discriminated against and this may risk potential reputational damage. Additionally, internal auditors and regulatory supervisors expect to understand the precise basis of solvency reporting.

Transparency and understanding are important parts of the European Commission's Artificial Intelligence Act which is currently being finalised. Without an explainable solution, organisations may face fines or the suspension of their ML deployment programmes.

Optimisation

Model outcome data is extremely large and should be stored in the data staging area, from where it can be recycled as additional data variable inputs.

Many organisations struggle to recycle outcome data, but this is critical to enable their further evolution from predictive insight ('what will happen') to more dynamic prescriptive insights ('what should be done to get the best possible outcomes')



HOW WE CAN HELP

✓ **Model monitoring and validation**

We provide a range of model monitoring and validation services which can be accessed either through cloud-based platforms or through specialist consultancy support. These include:

- Specialist tools and configurable templates
- Automated monitoring and validation applications
- Automated performance reporting and visualisation
- Automated alerts and recommended actions
- Model archiving to meet internal audit and supervisory requirements

✓ **Explainability**

Our clear and proven methodology assesses the relative contribution and level of influence of each variable to the overall score – enabling organisations to demonstrate transparency and fair treatment to auditors, regulators and customers alike.

✓ **Optimisation**

Our specialist optimisation services help to evaluate the best value trade-off between potentially competing priorities such as risk/revenue/cost.

Our decisioning and analytics platforms enable the capture of model outcomes which can be used to identify and define emerging business challenges.

Outcome data can be held in the data and analytics platform staging area so it can be reviewed against market data, macro-economic forecast data, and trended bureau data to support:

- Lost opportunity analysis of risk behaviours of rejected cases ('reject inference analysis')
- Analysis of portfolio risk concentrations by sector, segment or geographical region
- Peer group comparison of customer share of wallet/delinquencies
- Development of segment-specific models and scorecards

In this way, we help create a cycle of improvement and best practice.

Conclusion: The journey to maturity

Todays forward looking organisations are at differing stages of maturity in terms of their adoption of ML.

ML and AI have seemingly unlimited potential to create value.

As organisations progress towards higher levels of maturity, developing and deploying more and more models, realising the full potential is increasingly dependent upon more than just analytics.

> "Alignment, collaboration and orchestration increasingly become the key enablers of strategic risk agility."

The need for collaboration and alignment can be seen with regard to customer management operations and regulatory risk.

Customer management operational teams will use early warning models at an individual customer level to identify the first signs of vulnerability and intervene to stop cases from moving into collections.

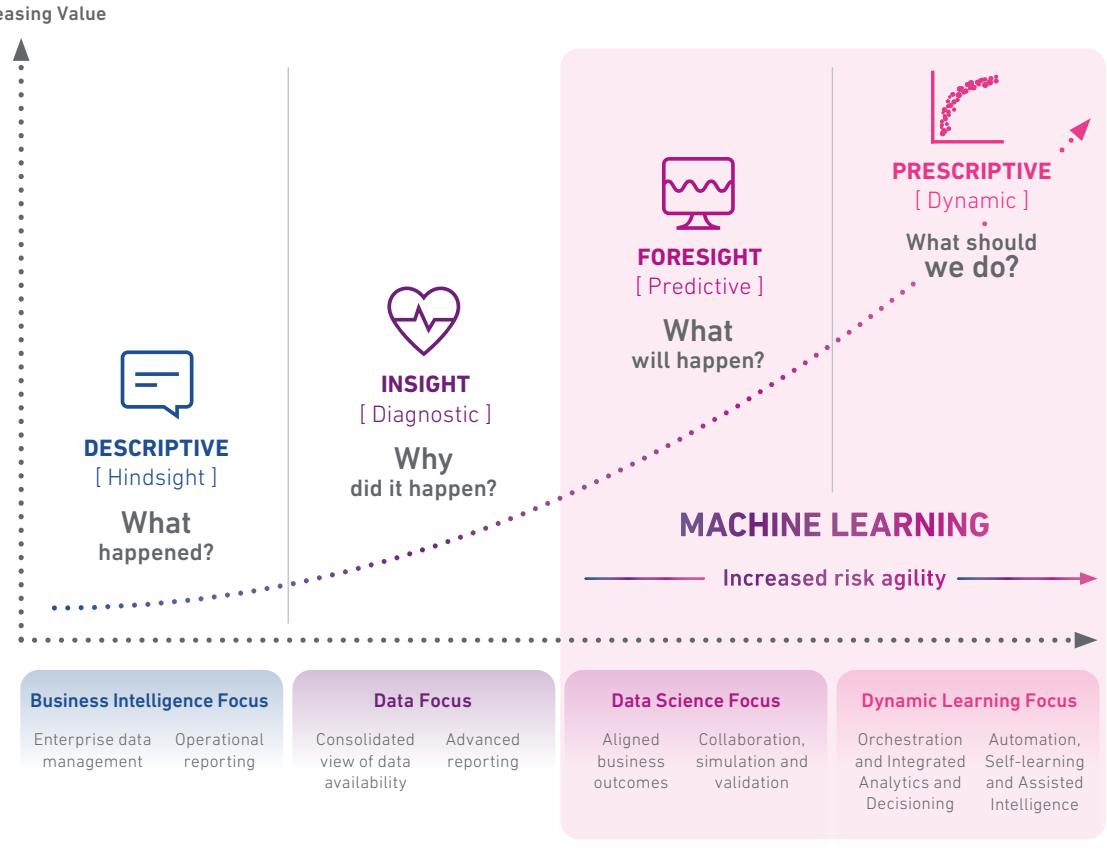
At the same time, regulatory risk teams use their own models at a portfolio level to predict levels of non-performing loans and their impact on expected credit losses and financial provisions.

Being able to ensure the alignment of data, models, reporting and visualisation across customer, segment and portfolio views will help ensure maximum visibility and with it provide the organisation with maximum agility and responsiveness.

The more models that are developed makes orchestration increasingly critical as multiple models are combined within specific data-to-insight-to-action pathways.

For example, within customer application and onboarding processes, there is a need to combine risk-based models that assess credit, affordability, identity and fraud, with value-based models that assess both the customer's potential and the propensity to purchase additional products and services.

In these instances, maximising efficiency whilst simultaneously minimising cost will require identifying and then orchestrating the optimal model application.



Wherever you are on the journey let Experian help you leverage the power of ML

With our unique blend of data, analytics, technology and expertise we can help bring increased visibility across multiple dimensions of risk and optimise strategies that deliver multiple business outcomes.

To find out more, get in touch.

Please contact your local Experian office or visit the Experian Academy website.

Data | Analytics | Technology | Expertise