An Experian Case Study

Trusso: Experian's categorisation engine

Introduction

Experian's DataLabs were established in 2010. Comprising of expert data scientists they focus on new data, advanced analytics and innovative techniques to solve complex client challenges.

One such project set out to investigate how new data sources, such as transactional data seen through Open Banking APIs, could enrich insight and equip businesses to make better decisions. For example, helping lenders better assess affordability and credit worthiness.

The DataLabs team believed they could achieve this by accessing and understanding greater detail around an individual's income and expenditure. They hoped that, as well as improving affordability and credit-risk decisions, this better understood data would offer opportunities across the entire customer lifecycle. Specifically, that it could be modelled to provide better-tailored strategies that benefit businesses and society.

"The project far surpassed our expectations. Our categorisation engine, Trusso, has performed with 90%+ accuracy on unseen data. We hope this paper gives an insight into our approach, methodology and findings.

Throughout development we worked to ensure decision-making was fair, accurate, customer centric and transparent. We use this approach – known as FACT – for every DataLabs project. We believe it's essential for work in machine learning and artificial intelligence to follow this principle.

Experian believe that data can be used to make better decisions and, in turn, transform lives. Similarly, we believe data science has huge potential for all, and we're investing in new technologies to help our customers maximise every opportunity. Open Banking is one such opportunity."

Javier Campos Zabala
General Manager
Experian UK&I DataLabs

Watch Javier explain more about the development of Trusso

The challenge

Today's digital, data-fuelled economy means many businesses are struggling to extract the right level of insight to inform their operations and decision-making strategies.

At the same time, the UK's introduction of Open Banking and the Payment Services Directive 2 (PSD2), as well as the FCA's consultation into creditworthiness, means the way in which decisions are made, or should be made, is changing significantly.

Many business leaders are assessing whether Open Data has the potential to change and enhance lending processes we see today. To do this requires lenders to think differently, and model assessments they do today differently. This was the challenge that the DataLabs were looking to help with.

What we did

Research project overview:

We were challenged to find a solution whereby Banks and other Financial Service providers could maximise data, specifically Open Banking data and transactional data sharing, to better inform decisions. That could mean credit decisions or deciding which products best suit which customers throughout the lifecycle. However, in its raw state, transactional data just adds more complexity to the already challenging Big Data landscape.

It was immediately evident in our investigations that the ability to categorise the data would be essential. Categories that correlated to actual spending and income patterns, such as utility bills, or recreational spending, would make the data more meaningful.

What we did

We created a strategy using a small set of test data and gained enough evidence so that our approach would scale for larger datasets.

We planned to use machine learning to group expenses, reducing the overall flow of data and creating meaningful and usable categories which would give insight into financial behaviours.

At first we started with a rules-based approach but soon saw the added benefits that a self-learning machine based system could bring to the task.

We intended to use the data to provide a granular depth of insight into customer behaviours. The taxonomy required a certain level of granularity. Here we needed to ensure the level of detail presented back was enough to give the required insights.

Connecting the data was also important. As well as matching the data with our own Experian data, we combined it, updated it, completed it where needed and created a single customer view. We then enhanced it by categorising through similarities and aligned it to a clear taxonomy framework. The output of the project was Trusso. Experian's categorisation engine

> With a project like this there are three areas for consideration – price, quality and speed. But you can only pick two. If you want fast and cheap, you'll lose the quality.

Rules-based vs. artificial intelligence

Most industry products have used rules-based decision-making for years. But given the scale, variety and newness of data in this environment, we believe artificial intelligence and machine learning is a more appropriate architecture. Our initial designs and tests proved this.

To build a rules-based categorisation engine for bank transaction data, would take between 20,000 and 40,000 rules, which would need constantly monitoring and changing as the data updates. The system could fail following key changes, such as a new retailer opening or if the system gets data from a new bank that labels cash ATM transactions differently.

Trusso's ability to self-learn is what makes the tool valuable and future-proof.

Self-learning brings more granular insight

One of our objectives was to enhance decision-making using this new data, to facilitate fair and responsible lending that deliver better outcomes for people.

This is hugely advantageous for segments of society where income and expenditure is inconsistent or incomplete, such as thin-file applicants or the self-employed. A significant proportion of small businesses currently struggle to access credit, due the volume of small deposits that aren't easily identified without this level of data science. Our machine learning system can gauge, aggregate and categorise income to offer a much clearer view of the customer.

Adding insight from Experian data

During development we added Experian data into the mix. We connected it to our digital identity capability and created both a consumer and commercial view in the architecture.

As Trusso expands we can quickly integrate new APIs, to add even more granularity and depth through different data sources. More importantly, clients can add their own data too.

Combining data on this scale gives businesses a powerful platform for developing products and propositions and helps provide an infrastructure that can be hugely beneficial to helping understand and better use data.

Testing: what we saw

The system works by learning and for this we needed annotated data. The more data, the better the system performs, as our tests proved.

The development and use of algorithms within Trusso provides specific features from the data. Instead of creating rules, we let the system decide and learn. From here we were able to use this algorithm to take in new data, creating a sustainable model that's continually developing. While the tool was developed in response to Open Banking, its usefulness far surpasses that.

Trusso looks across the entire system to identify regularities in data. It's therefore hugely beneficial for detecting salaries and can categorise irregular transactions through self-learnt rules. This is useful for credit applications, and is a reliable way to verify application data.

The system has 94 categories for income and expenses. These are grouped to give an exceptionally granular figure, including main income sources, salary, bonus, refunds, benefits, pensions, income protections and more. We found that this gave us the right combination between being granula, picking up signals in the data and accuracy. Any further grouping would prevent us from receiving a lot of the valuable and potential signals from the data.

Many UK Financial Services firms have variations in how expenses are grouped, while insurers may gather different application data. The engine allows for this, offering the flexibility of different rules for the company that's using it.

The results

Our tests have returned accuracy rates of more than 90% on data that has never been seen before and the figure is still increasing.

It's important to note that the theoretical limit is not 100% due to the lack of information and inherent ambiguity of certain transactions. Also, we're only ever reporting the accuracy of unseen data and not training data. Training data is skewed by the machine's memory, so it is not a true test of how well the Trusso would perform when presented with a new dataset. This accuracy is a fundamental business requirement that we have achieved.

This new form of data analytics enables any adopter to not only calculate risk better than ever before, but also to see signals from the data that can benefit the entire lifecycle of a customer. For example assisting through prior knowledge if a customer is nearing financial difficulty, or in the areas of upsell, crossell and retention.

This is hugely rewarding given the need to ensure lending is not only based on affordability criteria, but is affordable – always. It also makes the ideal of personalising an entire customer journey a reality, as a firm can make informed decisions throughout the lifetime of a relationship and use this to bring much more value to the relationship.

In this highly competitive marketplace, it enables users to have a layer of personal insight that can inform better partnerships and alliances across industries, underpinning the much needed 'value exchange' between a business and a customer. For example, the taxonomy used can show where a customer is spending and pick up on these patterns to identify where value can be given. For this reason, it would be especially beneficial to businesses offering special incentives.

Trusso is now operational in the UK and ready to accommodate the new data being shared the Open Banking and PSD2. The inherent flexibility of the system and its advanced data science mean that it can be deployed in a matter of weeks to enable implementation in a new country.

Trusso is very simple to integrate, taking a matter of weeks to build and deploy.

Flexibility and control

The engine is flexible and supports multiple taxonomies, allowing for different categories depending on what you want to do.

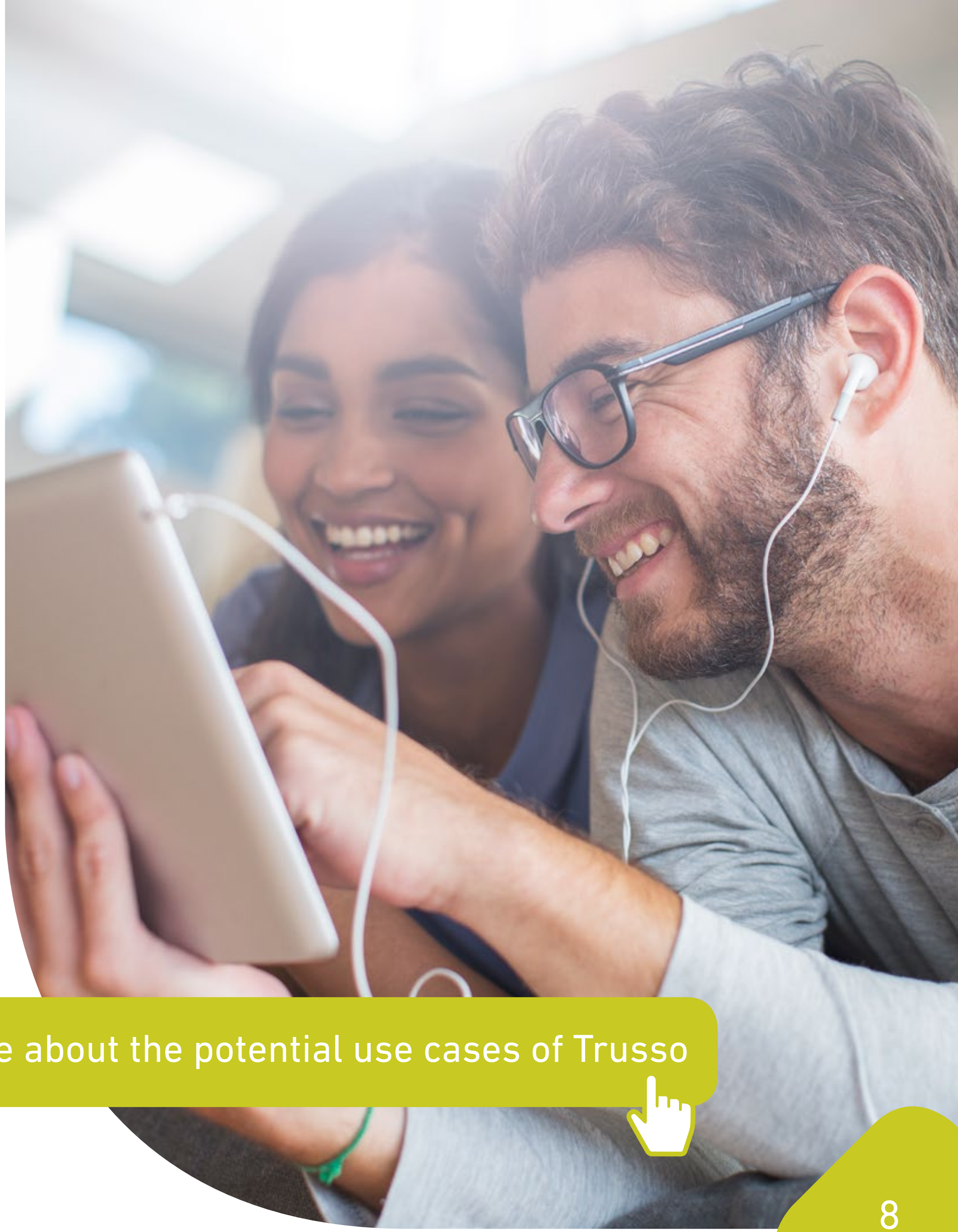
For example, it can be used as a precursor for assessing credit risk. If you're interested in marketing, it can be used to inform better customer targeting and management.

The way we've approached categorisation allows for multiple uses. It can be used to signal a change, providing better insight for credit risk or fraud teams for example. Or it could also be used to enhance cross sell opportunities.

If someone rings a call centre, the agent can receive instant information that supports the call and helps inform, and direct, the conversation.

Rules-based solutions would not support this, due to the level of manual intervention required to align rules to the client requirements. Trusso learns from itself instantly, meaning they are no longer a barrier to deployment.

The use cases of Trusso extend beyond Open Banking. It provides an approach and architecture that solves many Big Data problems.



Read more about the potential use cases of Trusso

What we believe

We're moving to a place where investments in data analytics are top of the agenda for many businesses. In addition, businesses no longer have months in which to deliver new initiatives, and react and respond to disruption and influences. The market is reaching saturation point and we believe any development today needs to be scalable for the future. The art of data science, and tools such as Trusso, enable a quick response and are scalable, at speed.

Data is owned by the individual; the GDPR reinforces this. People expect a clear and valuable exchange when it comes to interacting with businesses. That value can be determined through data, so long as it is understood and used appropriately. Being able to categorise data, and enhance it, will help meet this need.

Moving forward, the concept of Big Data will only grow. The Internet of Things, connected devices and many more phenomena have long been reported as revolutionary, bringing in actual behavioural insight for the first time. Transactional data is no different. Our engine is scalable for all types of data and the architecture of this means we can continue to develop new propositions and potential for extracting insight from data through machine learning.

Conclusion

The potential and business value for categorisation through Trusso is enormous, especially in the context of credit risk and affordability, customer churn, retention or growth.

Today there is a need for all businesses to compete in a real time environment at a more sophisticated level. This means moving to a place where you can decide, in seconds, whether to offer a loan and what type of loan it should be. This is a massive change from where the market is today.

Machine learning is often deemed as a black box, offering quick and easy solutions to those who know little or nothing of the inner workings. So while the results are transformational, we believe they should be applied with discretion.

Machine learning offers a much more granular approach for the cases which require more accuracy than classical models.

However, where classical models suffice, we would advocate these be used in place of machine learning of this kind.

We are in a data insight environment. Many are not only catching up with society, but trying to innovate, balance risk and reward and compete too. These conflicting priorities mean fraud is going undetected and engagement is declining. This is due to the lack of a clear 'happy' customer journey.

The concept of machine learning and artificial intelligence has long been a trend, but this trend is not going away. In fact, it's a solution that responds to trends offering an opportunity to accelerate business models and, ultimately, better serve businesses and communities.

The added benefit is how, if used correctly, it can enhance every Key Perfomance Indicator (KPI) too.

About the DataLabs

Experian DataLabs helps businesses solve strategic marketing and risk-management problems through an advanced data analysis process and research and development. Focused on innovating new data sources with an emphasis on financial services, telecommunications and healthcare, DataLabs helps deliver:

- Increased profitability
- Optimisation of data assets
- Controlled financial risk
- Regulatory compliance

Experian DataLabs is staffed with a multidisciplinary group of data scientists with Ph.D.s and applied research practitioners with expertise in advanced analytics and machine learning, as well as other advanced statistical methods. Experts in applying cutting-edge data science to real-world business situations.

> "Experian DataLabs provides a safe and secure environment in which to partner with our clients, enabling breakthrough data experimentation and innovation."

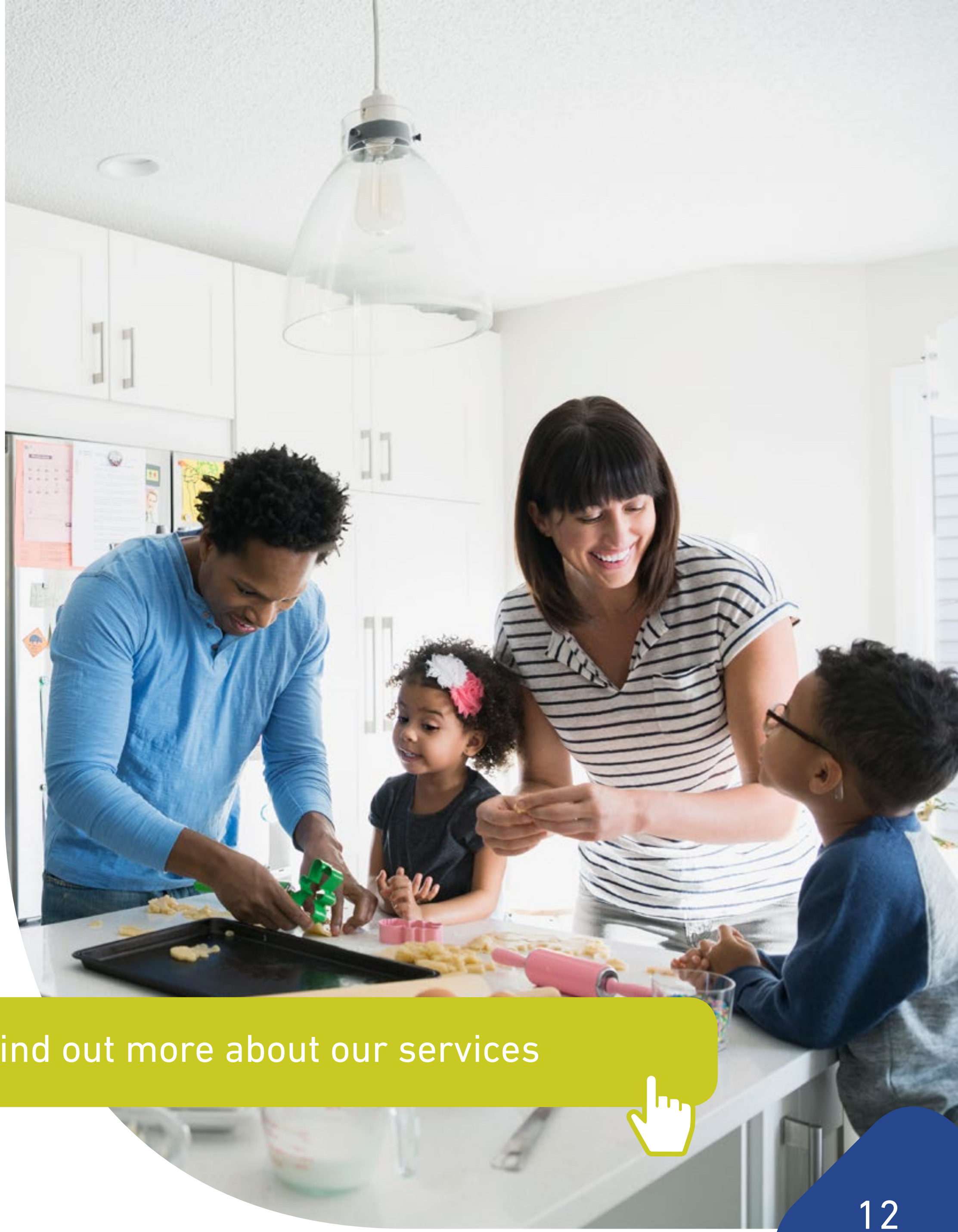
Eric Haller
Executive Vice President
Experian DataLabs

Learn more about DataLabs' other global projects

About Experian

Experian unlocks the power of data to create opportunities for consumers, businesses and society. At life's big moments, from buying a home or car, to sending a child to college, to growing a business exponentially by connecting it with new customers, we empower consumers and our clients to manage their data with confidence so they can maximise every opportunity.

We gather, analyse and process data in ways others can't. We help individuals take financial control and access financial services, businesses make smarter decisions and thrive, lenders lend more responsibly and organisations prevent identity fraud and crime.



Find out more about our services

What is FACT?

Fairness, Accuracy, Customer, Transparency

Experian DataLabs helps businesses solve strategic problems across marketing and risk-management, through an advanced data analysis process and research and development

Fairness

We make sure that any decision isn't discriminative by removing any bias from the data. E.g. race or gender.

Accuracy

Accuracy is achieved by identifying, recording and articulating any sources of error or uncertainty throughout the algorithm, including its data sources. Greater accuracy leads to better outcomes.

Customer

The customer needs to be at the heart of any task, any project or any analytical approach, and any output or objective needs to be centred around what's best for them.

Transparency

Transparency is critical. How the data is processed and used to reach a decision needs to be easily communicable.

New technology brings added complexity: how can we ensure machine learning and artificial intelligence are used responsibly? Fairness, accuracy, transparency and the need to put the customers' best interests first, are pivotal to any task undertaken at DataLabs.

It's also important to help third parties understand the algorithm's behaviour and allow them to monitor, check and review it. The provision of detailed documentation, technically suitable APIs and permissive terms of use will be integral to this.

Glossary of terms

Categorisation

Categorisation systems automatically classify all expenses and income transactions from multiple bank accounts (including current accounts, debit and credit cards), into a pre-defined group of categories, or taxonomy.

Rules-based system

A system built on predetermined rules, typically requiring between 20,000 and 40,000 rules to be manually coded. In our view, these satisfy certain needs and will continue to be a valuable tool. However, rules-based systems don't offer the scalability or agility required to enrich insight and inform strategies.

Artificial intelligence (AI)/machine learning-based systems

Machine learning algorithms such as Trusso can work out how to perform important tasks by generalising from examples. They remove the need for manual rules, as the machine learns patterns within the input data. For this reason, such systems are often feasible and cost-effective where manual programming isn't. As more data becomes available, machine learning can tackle more ambitious problems.

Taxonomy

The Experian Taxonomy has been designed to be configured to any client's requirements. The current default taxonomy has 18 branches covering income and expenses (11 for expenses, 6 for income), plus 94 categories of transactions. The flexibility of the taxonomy is probably one of the most important aspects of a categorisation engine, allowing it to be grouped in whatever way is best for that use case. Moving forwards it is likely the taxonomy will change as we adjust to the learnings from new data.

Generalisability and accuracy

It's very important to have the right accuracy and on new unseen datasets. As data continually changes, an AI/ML-based engine will outperform those based on manual rules over time. Accuracy on unseen data allows us to test the generalisation capability of the model, as it will be applied to characteristics, signals and attributes in the data that we haven't seen. A rules-based approach can score very highly on 'trained' data, yet it fails when presented with 'untrained' data containing signals that have not been seen before.

Accuracy: theoretical limit and trends

We cannot expect the machine learning engine to be better than a human, since the training data itself is created by humans. There will always be transactions that are ambiguous, meaning they can fall into multiple categories. This is especially true when the taxonomy is more granular. But the benefits of a granular taxonomy outweigh this issue by far. Due to this ambiguity, the engine will not be able to have 100% accuracy.

The theoretical limit and upper bound depends on a given taxonomy and the dataset under observation. Studies on a very granular taxonomy have show that where the same transactions are assessed using agreed labels, consensus is achieved on roughly 93-94% of transaction.

Registered office address:
The Sir John Peace Building, Experian Way,
NG2 Business Park, Nottingham, NG80 1ZZ

E: gtmcontactus@experian.com
www.experian.co.uk/b2b