Designing for Trust with Machine Learning

Fabien Girardin, Pablo Fleurquin
BBVA Data & Analytics, Avenida de Burgos, 16D 28036 Madrid, Spain
fabien.girardin@bbvadata.com, pablo.fleurquin@bbvadata.com

Abstract

This is a proposal for a presentation on the relation between Machine Learning and design for trust at the Designing the User Experience of Artificial Intelligence symposium as part of the 2018 AAAI Spring Symposium Series in Palo Alto, CA. Trust is at the bedrock of our human social system. Historically, the financial businesses have been based on how it could trust customers, and not the other way around. Today customers request — in addition to competence, security and lending capability — honesty, legibility, transparency and other key attributes of the trust relationship with a data-driven bank. We will share our experiments and approaches that use Machine Learning techniques to tackle mistrust and foster a trustworthy relation with our customers.

Introduction

Fabien Girardin is Co-CEO at BBVA Data & Analytics, a center of excellence in financial data analysis that aims at revolutionizing the banking industry in the domains of marketing intelligence, customer advisory, risk, fraud and the automation of financial processes. With a broad spectrum of interdisciplinary skills, he guides teams in transforming algorithmic research and experiments into value propositions, services, products and experiences that are future forward.

Pablo Fleurquin is Data Scientist at BBVA Data & Analytics with extensive experience in describing, analyzing and modelling the delay dynamics of a paradigmatic socio-technical complex system such as the air-transportation system. He uses his knowledge in Complex Network Theory, Graph Analytics and Machine Learning to develop online credit card fraud analytics, risk scoring solutions and pricing strategies.

This paper reports on our investigation and experiments that explore how the specific design of Machine Learning algorithms can consolidate trust in financial services. This work aims at orienting today how people experience banking in the near future.

An Evolution of Trust

Trust is part of a social contract with both rational and emotional bonds. Trust cannot be delivered, but actions can be taken in order to enrich it. For instance, financial businesses are based on their capacity to measure risk to grant a loan or accept a transaction. Historically, quality, transparency and altruism was demanded on the side of the customer. In consequence, a bank is often perceived as a partner people need to live with, but that are prone to mislead, provoke unfair situations and take advantage of opaque processes. That situation is changing with regulators and society, in various parts of the world, demanding openness for both protection of personal data and therefore breaking bank's monopoly in measuring risk.

Nowadays the increasing amount of digital footprint of bank customers provide with a much deeper vision to measure risk and opening new means to build trust. New analytical capacities like Machine Learning allow to transform these new datasets into personalized experiences, customized advisory with accurate forecasts, increased access to loans with less risk, as well as automated interactions.

Those technological opportunities also create design challenge that may drive mistrust between banks and their customers. The practice of data science must carefully resolve an increasing amount of dysfunctional solutions based on partial data or in bad quality data. Importantly, Machine Learning errors have totally different implications depending on the domain: the consequences are very different if we are recommending a financial product, a movie or helping with illness diagnose. Those solutions have the potential to erode trust and disengage customers, besides posing a risk proportional to the kind of service provided. A lot of research interest has been put recently on adversarial examples. These are subtle and unnoticeable changes to model inputs that an attacker intentional designs to cause

the algorithm to make a mistake. For instance, in the facial-recognition field where the industry and government intelligence agencies have put a lot of effort, a recent paper has shown how by changing a small part of the image is enough to make you a different person in a machine's eyes (Sharif et al. 2017). Another research group has also shown that street-signs recognition algorithms for self-driving cars are also prone to adversarial examples. Subtle changes that a human will recognize can make an algorithm confuse a stop sign with a speed limit one (Papernot et al. 2017). In addition, discrimination like unfair access to societal goods is becoming pervasive and has reinforced the threat. We have to highlight that these technological threats, as opposed to adversarial examples, happen without any explicit wrongdoing in Machine Learning modelling. Two of the main reasons behind such a pervasive problem are sample size disparity and encoded human biases in data. The former is easy to grasp, basically minority groups are by definition under-represented in data sample, which leads to higher error rates on these groups. The latter, is part of the data and in most cases is indistinguishable from it. Biases come in many flavours: demographic, geographic, behavioural and temporal biases. Examples are becoming ubiquitous such as 2013 Ally Financial 98M US$ suit on auto-loan discrimination (McDonald and Rojc 2014). In this particular case, the Consumer Financial Protection Bureau's (CFPB) used an algorithm to infer a borrower's race based. Other border-line use of technology is in recidivism models such as the LSI-R in the United States (Whiteacre 2006). These solutions help the judicial system to assess the danger posed by each convict. A work by Caliskan et. al. 2017 showed how pre-existing biases and stereotypes permeate semantically derived word associations models. It is clear, though, that algorithms inherit human biases, that pervade historical data, and the situation is even worse when these are camouflaged into a black-box model.

Up to this point, we believe any data-driven organization like a bank we must be transparent and responsible through their decision-making process, being it algorithmically driven or not. Hence, they must detect and address potential problems to enrich a trustful relation with their customers. Our work in that domain explores the foundations of trust from a Machine Learning perspective with the basic attributes of fairness and transparency.

Fairness is always the result of a comparative process (Xia et al. 2004). This can be twofold; as a comparative process with a past personal situation or a comparative process with another person independently of time. For example, in the former case, we can consider a price increase in a certain product, given incomplete market information, as unfair. In this, anticipating the buyer discrepancies and the transparency of the vendor explaining why price has increase can reduce the sensation of unfairness. In the latter case, we base our fairness assumptions by comparing to others. Things are more intricate, because one must address, subjectively, how alike one is to the comparative others. If there is a price reduction in a certain product for people considered as peers, odds are that the comparison will provoke an unfair situation. A good example of it was the uproar that took place with Amazon dynamic pricing model when people realized that the model had charged some people more than others (Weisstein et al. 2013). Unfairness of the second type can be explicitly solved in the feature selection phase (Grgic-Hlaca et al. 2018) or including fairness metrics as another component of the algorithm development (our experiments 2 & 3).

In addition, transparency also known as Machine Learning interpretability is a key part of the toolset to tackle mistrust in algorithmic decision-making processes. It can be used to promote fairness of the first and second type, and moreover pervade the organizational culture with ethical responsibility. As the great 20th-century physicist Richard Feynman puts it: "if you cannot explain something in simple terms, you don't understand it". This maxima that is so accepted in the hard sciences, it is not that extended in Data Science. It implies a bidirectional association between explainability and understandability, which ultimately oppose transparency against blackbox-ness. It should be noted though, that black-box algorithms are not exclusively those of a non-linear nature; high dimensional and heavily tuned Generalized Linear Models can be also vastly opaque (Lipton 2016). Fortunately, interpretability frameworks clear the way to take-apart the machine and explain its pieces (our experiment 1) (Ribeiro et al. 2016; Lakkaraju et al. 2017).

An Evolution of Automation

Automation in the banking industry has come a long way since the 1970s with innovations like the Automated Teller Machine (ATM) and the Electronic Fund Transfer at Point of Sale (EFTPOS) (Consoli 2008). Automation is in the DNA of such an information driven industry. In the last years with the advent of cheap distributed databases, cloud services and computational power automation pivoted to enrich decisions algorithmically by incorporating vast and varied new data sources. Nowadays, many banks follow a digital agenda focusing on sales automation. By doing so, personalized offers reach customers at the right moment, and, in addition, automating servicing 'Do it Yourself' experiences allow for huge cost reductions on mature high margin products. Also, data-driven banks employ Machine Learning to perform more fine-grained assessment of risks and provides customized advisory.

According to McKinsey Global Institute Report (2016) Machine Learning is having a significant impact on retail banking, especially on improved forecasting and predictive

analytics boosting a radical customer personalization approach (Henke et al. 2016). Nevertheless, the evolution of automation should come along with that of trust, but this coevolution is far from clear. According to an Accenture poll 87% of US consumers plan to use bank branches because of greater added value and in-person trustworthiness (Accenture 2016). Still, in general, the most valuable channel is online but not precisely because of trust as it is the reason behind branch channel value. Automation might move from traditional transactional interactions to a meaningful "relational" interaction. In an increasingly digital era, consumers are looking for experiences rather than merely servicing; a world where banks come to customers rather than customers go to the bank. Therefore, the interplay between automation and customer experience should come along together with trust, and this area is where we are putting our research efforts: how the design of automation together with Machine Learning can create trustworthy relationships with our customers.

Experiments on Trust and Machine Learning

We are currently conducting experiments that aim at understanding techniques to design for trust with Machine Learning

- Experiment 1 is about interpretability and trust in credit risk scoring: Algorithmic transparency is openness about the purpose, structure and underlying actions of the algorithms used to search for, process and decision making. This experiment explores one way of making a black-box algorithm transparent using LIME (Ribeiro et al. 2016) as an interpretability framework. By implementing this framework we can answer customer questions such as: why I have been rejected? Not only for the customer but also for the financial regulator which opens the possibility to use more sophisticated non-linear models. As well as helping risk analysts on the model development process.
- Experiment 2 is about learning to bid in real time using a fair strategy: An approach on dynamic pricing that uses Reinforcement Learning (RL) (Sutton and Barto 1998) to keep a balance between revenue and fairness. This work helps maximize revenues while taking into account fairness and equity that prevent a negative customer perception of unfair price differences that can destroy a trustful relation. We demonstrate that RL provides two main features supporting fairness in dynamic pricing: on the one hand it is able to learn from recent experience adapting the prices policy to complex market dynamics; on the other hand RL can include a trade off between short and long-term objectives, integrating fairness into the model's core. Specifically Q-learning is used to provide a simple way for agents to learn sequentially by trial and error (Watkins and Dayan 1992). In the context of our experiment it is used to, for each action performed by an agent, modify the state of the environment (related to fairness) while providing a reward (the price bid).
- Experiment 3 explores a fair approach on Recommender Systems (RS): While RS aim to provide an appealing list of items to users, most algorithms suffer from a bias in the recommendation towards popular items. As a consequence, the recommended list often goes away from the true interest of users. On the other hand, less popular, long-tail items are desirable for recommendations because of their novel and diverse character. In this experiment, we explore the concept of fairness in recommender systems, so that all items have the same chance to be presented to users. Two techniques that allow keeping a balance between popular and niche products in the recommendation are introduced. A new loss function that it is explicitly designed to deal with missing information, forbids a predicted zero preference to unseen products. This makes every product available in the recommendation. Second, a popularity-scaling factor is included in the loss function distributing the recommendation itself in a better way.

Conclusions

Trust is a complex term with multiple dimensions investigated in psychology, sociology, economics, information systems and even philosophy. From a Machine Learning perspective, we realize to only grasp the tip of the iceberg. With the new wave of Machine Learning solutions, value is created with an accumulation of touch points that feed algorithms with behavioural data. Technology can provide attributes to build trust like competence, quality, simplicity, and convenience. We have seen that Machine Learning technique can help contribute to further experiences of trust like transparency and fairness. We believe that trust is built through the intensifying relations, feedback loops, virtuous cycles, 'data network effects', and the capacity to understand and react on customer's intentions, emotions, and behaviours.

We believe that models are not sanitized abstractions of reality; on the contrary, explicitly or not, they are being created with our biases and unfair judgments. These must

not be seen solely as profit seeking machines, because the choices they made in the end are fundamentally moral.

In addition, we are exploring ways to include fairness and transparency as central elements of model development, that eventually will foster a trustful relation with bank customers. We have learned one way of making opaque algorithms transparent positioning us one step ahead of the new regulatory demand which comes into force next year under the European Union General Data Protection Regulation (GDPR). Using model interpretability, we can fulfil the regulatory "right to explanation" and give feedback to customers on the decisions that affect them, as well as help data scientists and analysts on the process of training and assessing models. Regarding fairness, we are gathering empirical evidence that Reinforcement Learning is a model capable of learning revenue maximization while providing a more egalitarian dynamic pricing strategy between groups of customers. Concerning recommender systems, we developed a way of effectively dealing with the extended bias in recommendation towards popular items. Avoiding this bias means new responsible ways for the banking industry to increase its sales and profits by potentially selling in a vast and unexplored market.

References

Accenture Consulting. 2016. North America Consumer Digital Banking Survey. Banking on Value: rewards, robo-advice and relevance, Technical Report, Accenture.

Caliskan, A.; Bryson, J.J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334), pp.183-186.

Consoli, D. 2008. Systems of Innovation and Industry Evolution: The case of Retail Banking in the UK. *Industry and Innovation* 15(6), pp.579-600.

Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P. and Weller, A., 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. Max Planck Institut Informatik

Henke, N.; Bughin, J.; Chui, M.; Manyika, J.; Saleh, T.; Wiseman, B.; and Sethupathy, G. 2016. The Age of Analytics: Competing in a Data-driven World, Technical Report, McKinsey Global Institute.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J., 2017. Interpretable & Explorable Approximations of Black Box Models. *arXiv preprint arXiv:1707.01154.*

Lipton, Z.C. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490.*

McDonald, K.M.; and Rojc, K.J. 2014. Automotive Finance Regulation: Warning Lights Flashing. *Bus. Law.,* 70, p.617.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; and Swami, A. 2017. Practical Black-box Attacks Against Machine Learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (pp. 506-519). ACM.

Ribeiro, M.T.; Singh, S.; and Guestrin, C. 2016, August. Why should I Trust You?: Explaining the Predictions of any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M.K. 2017. Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition. *arXiv preprint arXiv:1801.00349.*

Sutton, R.S.; and Barto, A.G. 1998. *Reinforcement learning: An introduction* (Vol. 1, No. 1). Cambridge: MIT press.

Watkins, C.J.; and Dayan, P. 1992. Q-learning. *Machine learning* 8(3-4), pp.279-292.

Weisstein, F.L.; Monroe, K.B.; and Kukar-Kinney, M. 2013. Effects of Price Framing on Consumers' Perceptions of Online Dynamic Pricing Practices. *Journal of the Academy of Marketing Science* 41(5), pp.501-514.

Whiteacre, K.W. 2006. Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review* 17(3), pp.330-342.

Xia, L.; Monroe, K.B.; and Cox, J.L. 2004. The price is unfair! A Conceptual Framework of Price Fairness Perceptions. *Journal of marketing* 68(4), pp.1-15.