Exploring the Role of Domain Experts in Characterizing and Mitigating Machine Learning Errors

Pavel Hoogland[1], **Margje Schuur**[2], **Piet van Hoffen**[2], **Jasper van Vliet**[2], **Oana Inel**[1], **Jie Yang**[1]
[1] Delft University of Technology
[2] Inspectie Leefomgeving en Transport, Dutch Ministry of Infrastructure and Water Management
P.Hoogland@student.tudelft.nl, {O.Inel, J.Yang-3}@tudelft.nl, {margje.schuur, Piet.van.Hoffen, jasper.van.vliet}@ilent.nl

Abstract

In the use of machine learning systems, end-users' trust can often be hard to attain as many state-of-the-art systems operate as black-boxes. Errors produced by these systems, without further explanation as to why the decisions are made, will deteriorate trust. This effect is especially strong when these erroneous decisions are generated with a high confidence. This paper presents a human-in-the-loop methodology to characterize and mitigate high-confidence errors by engaging domain experts through a series of interaction sessions. We study the problem in the context of Road and Transportation law violations, by engaging inspectors in day-in-the-life and in-house interview sessions. We show that by bridging the knowledge gap between domain experts and data scientists through these iterative expert sessions, we can improve the model predictions and achieve increased user trust.

Introduction

The problem of a disconnect between *data scientists* and *domain experts* is often present when using machine learning (ML) systems (Viaene 2013). This disconnect can be present on different levels, i.e., the concept or the process (how-to) level (Mao et al. 2019; Convertino et al. 2008, 2009) and can lead to decreased *domain experts* trust in the system. Providing model outcomes without insights into why the predictions were made, could harm user trust even further and inhibit user-developer interactions. This is particularly the case for erroneous outcomes. Thus, understanding why certain predictions are made, is key to user engagement, system adoption, and sustainability (Sousa, Lamas, and Dias 2014).

Errors produced by machine learning systems fall into two broad categories. Errors near the decision boundaries of a model are more understandable as they are caused by the inherent variances within the data. However, when an erroneous decision is made far from the decision boundary it can hint at inherent issues with the data used to train the model, whether it be a lack or under-representation of data points. These errors are produced with a high model confidence, namely the High-Confidence Errors (HCEs). These should be avoided and kept to a minimum to preserve trust.

Data selection and feature construction can be seen as the main crux of classical machine learning. Previous work (Fails and Olsen 2003) has shown that for certain applications, a continuously interactive way of machine learning can lead to improvements. A study into development tools for statistical ML (Patel et al. 2008) has shown that there is a need for an exploratory and iterative process in the process of data, and feature selection. In the natural language processing domain, (Park et al. 2021) proposed interactive tools that enable sharing domain knowledge through domain concept extraction and label justifications. Nonetheless, the role of the *domain experts* themselves is often overlooked in the development of trustworthy machine learning systems.

This paper explores the effect of iterative experts' engagement in a series of interaction sessions on understanding the requirements, challenges, and characterizing system errors. We show that these sessions are a promising method to facilitate model development and build trust with the model.

Methodology

We start by describing our proposed *Iterative Expert Sessions* to bridge the knowledge gap between *domain experts* and *data scientists*. It consists of three types of sessions, namely *exploratory*, *in-depth*, and *analysis*, aiming to close the knowledge gap and help data scientists improve the model in an iterative manner. A graphical summary of the sessions is displayed in Figure 1.

Exploratory sessions aim to get a deeper understanding of the working practice of the *domain experts*; learn which are the indicators or features that experts use to ground their decisions on, and how they compare to how the machine learning model is built. The driving question of these sessions is *"Where do the domain experts put their focus when making their decision?"*. To assess this, we propose either a day-in-the-life style session where the *data scientists* join the experts in their practice, or an interview setting where *data scientists* choose topics of expertise for which they need more insight. In this session, *domain experts* provide *data scientists* with insights in their reasoning, to help them better understand what features and data are essential for predictions.

In-Depth sessions aim to focus on those points from the exploratory session(s) that require further study. The focus lies on understanding what data is useful for the model and how much it contributes to the prediction – extracting and ranking features that are vital to the model's goal. The driv-

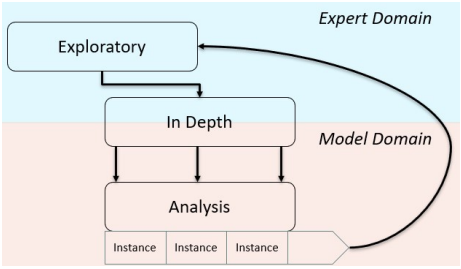

Figure 1: Graphical representation of Expert Sessions

.

ing question of this session is *"How does the available data contribute to the classification goal?"*. Answering this question sometimes requires further elaboration of information gathered from *domain experts* during the previous session. The in-depth sessions are performed as interviews, to offer space for discussions. The data scientists choose those areas in the data that they still have doubts in and ask from the *domain experts* to assess their relevance to the classification.

Analysis sessions serve two purposes: (1) present experts with actual use case data instances and model classifications and (2) provide hands-on experience with the model. By studying experts decisions and model classifications on actual data instances, we can better compare the model performance with how experts make decisions in real-world scenarios. Furthermore, we want to see how much the experts trust these model classifications. The driving question of this session is *"Do the domain experts trust the model predictions?"*. By accompanying model predictions with appropriate interpretation methods (we use a SHAP value variable contribution analysis (Lundberg and Lee 2017)), we offer *domain experts* hands-on experience with the model. The analysis sessions are performed as interviews. The *data scientists* present pre-selected data instances for which experts' input is informative, such as HCEs. The *domain experts* are to first give their assessment and classification of the instance. Then the model prediction is revealed and experts discuss similarities or differences in their judgement.

Case Study

We conducted this study in collaboration with the Netherlands Human Environment and Transport Inspectorate (ILT)[1]. The case study consists of assessing the risk of non-compliance for transportation vehicles on the Dutch roadways. The aim is to help road inspectors (*i.e.*, *domain experts*) to identify vehicles that are potentially non-compliant with road and transportation laws.

Data scientists at ILT developed a random forest based classifier which provides a risk score for potentially non-compliant vehicles, and thus can help select vehicles for examination. The classifier is trained on historical inspection data of Dutch vehicles. The historical inspection data is combined with data from other data sources such as vehicle and company registration.

[1]Inspectie Leefomgeving en Transport (ILT), Ministerie van Infrastructuur en Waterstaat: https://www.ilent.nl/

An *exploratory session* was held with 7 *inspectors*, 4 *data scientists* and 2 discussion coordinators. During this session the *data scientists* proposed 3 areas of exploration to the *inspectors*: overload, rest and driving times and cabotage. The *inspectors* brainstormed to identify and rank the most important indicators, while collaboratively taking notes. At this stage, *data scientists* can also ask clarifying questions. An *in-depth interview* session was held with 4 *inspectors* and 4 *data scientists*. Inspectors were queried on the following areas of interest regarding model data: moment and location of inspection, cargo, and maintenance, vehicle-ownership and company structures. Finally, an *analysis session* was held with 3 *inspectors* and 4 *data scientists*. The inspectors were presented with 8 data instances, the feature contributions (SHAP plot) and outcome of the model: 5 HCEs and 3 correct predictions. An image of the vehicle was also provided. Their judgement was used to assess the model performance.

Results

In the *exploratory session*, the inspectors concluded that vehicle violations are more prone to happen for certain types of freight/vehicles and on certain periods or days. This lead to revisiting of the feature set to better represent these aspects. The *in-depth sessions* led to conclude that more information about the vehicle maintenance and overload is needed. The vehicle maintenance information can hint at problems with the vehicle signaling a risk of error, and vehicle overload information can hint at overload risk. After making these changes, we saw a slight model performance improvement, but also less overfitting on biased historical data. In the *analysis session*, we found that experts agreed with all shown model predictions, disregarding the correctness of the prediction. Thus, at least for these cases, the model's judgement closely mirrors that of the inspectors and the presented HCEs do not deteriorate trust, since they still do indicate a potential risk.

In addition, we found that on-site visual characteristics of vehicles are extremely insightful for inspectors, but difficult to capture in the data. Nevertheless, even with a lack of visual features, the model still provides strong support for inspectors in the decision-making process.

Conclusion

We show that our proposed iterative session model can bridge the knowledge gap between *data scientists* and *domain experts*, in the context of road and transportation law violations. The model improvements made during the *exploratory* and *in-depth* sessions provided the inspectors with a more coherent prediction. We observed that some HCEs, even though an error at the time of inspection, still reflect a nature of risk of the vehicle - which helps in maintaining user trust in the system. Future inspections of these can still identify irregularities, making it difficult to distinguish real errors from model errors due to the temporal aspect. As future work, we plan to improve the effectiveness of the model by using a hybrid approach consisting of a predicted risk score, as well as providing feature importance values and decision rules, based on what the model has learned.

Acknowledgements

We would like to thank and acknowledge the Road Transportation inspectors that participated together with us during the case study sessions. Gert Bosman, Willem van Dijk, Jan van der Laarse, Govert Pelkmans, Erik Rozenbrand, Bert van Voorthuizen, Björn Weekhout. We also want to thank Mark van der Ham, teamleader of ILT Roadtransport, for helping schedule the sessions and select the inspectors. Special thanks to Karima Azzouz and Jacques Niehof, who helped coordinate the first exploratory session. Finally thanks to Victor Ciulei and Paul Ozkohen who helped provide feedback and insights during the development of the sessions.

References

Convertino, G.; Mentis, H.; Rosson, M. B.; Slavkovic, A.; and Carroll, J. 2009. Supporting content and process common ground in computer-supported teamwork. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*, 2339–2348. doi: 10.1145/1518701.1519059.

Convertino, G.; Mentis, H. M.; Rosson, M. B.; Carroll, J. M.; Slavkovic, A.; and Ganoe, C. H. 2008. Articulating Common Ground in Cooperative Work: Content and Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 1637–1646. New York, NY, USA: Association for Computing Machinery. ISBN 9781605580111. doi:10.1145/1357054.1357310. URL https://doi.org/10.1145/1357054.1357310.

Fails, J. A.; and Olsen, D. R. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, 39–45. New York, NY, USA: Association for Computing Machinery. ISBN 1581135866. doi:10.1145/604045.604056. URL https://doi.org/10.1145/604045.604056.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Mao, Y.; Wang, D.; Muller, M.; Varshney, K. R.; Baldini, I.; Dugan, C.; and Mojsilović, A. 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proc. ACM Hum.-Comput. Interact.* 3(GROUP). doi:10.1145/3361118. URL https://doi.org/10.1145/3361118.

Park, S.; Wang, A. Y.; Kawas, B.; Liao, Q. V.; Piorkowski, D.; and Danilevsky, M. 2021. Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models. In *26th International Conference on Intelligent User Interfaces*, IUI '21, 585–596. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380171. doi:10.1145/3397481.3450637. URL https://doi.org/10.1145/3397481.3450637.

Patel, K.; Fogarty, J.; Landay, J. A.; and Harrison, B. 2008. Investigating Statistical Machine Learning as a Tool for Software Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 667–676. New York, NY, USA: Association for Computing Machinery. ISBN 9781605580111. doi:10.1145/1357054.1357160. URL https://doi.org/10.1145/1357054.1357160.

Sousa, S.; Lamas, D.; and Dias, P. 2014. A Model for Human-Computer Trust. In Zaphiris, P.; and Ioannou, A., eds., *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*, 128–137. Cham: Springer International Publishing. ISBN 978-3-319-07482-5.

Viaene, S. 2013. Data Scientists Aren't Domain Experts. *IT Professional* 15(6): 12–17. doi:10.1109/MITP.2013.93.