What would you ask your ML model?

Explainable AI chatbot

Michał Kuźba[1, 2], Przemysław Biecek[1,2]

[1]Faculty of Matchematics and Information Science, Warsaw University of Technology

[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw

Introduction

What human operator would like to ask the ML model? To answer this question we have created a conversational system designed to explain behaviour of Machine Learning models. For this experiment we have trained a random forest model that predicts odds of survival from sinking of Titanic.

For this model we have implemented a chatbot called *DrAnt*. People can talk to *DrAnt* about model to understand the rationale behind its predictions. Having corpus of 1000+ dialogues we analyse the most common types of questions that users would like to ask.

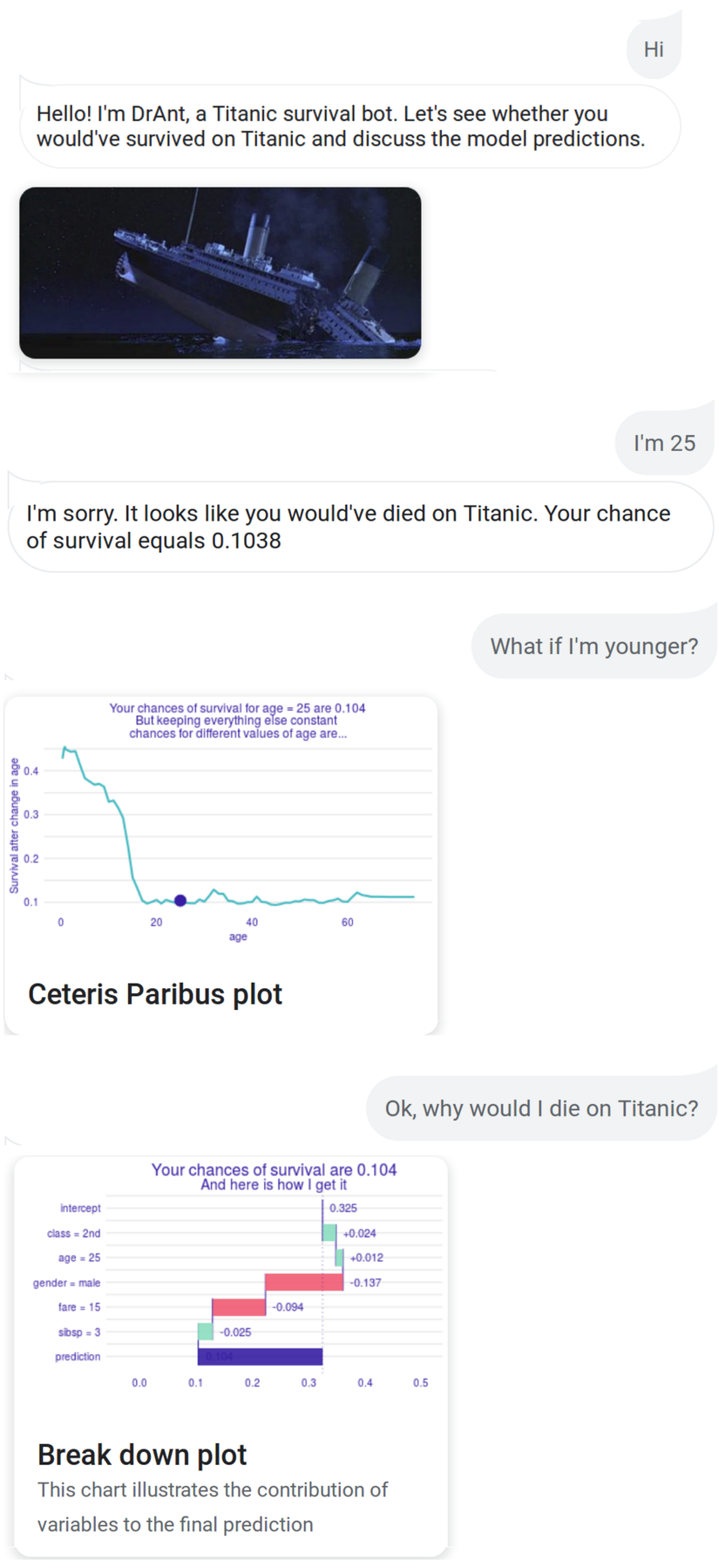
First, try it yourself!

scan & chat

XAI bot

Chatbot offers a conversation about the underlying Random Forest model trained on Titanic dataset. It understands and responds to several groups of queries:

- Supplying data about the passenger
- Inference
- Visual explanations from the DALEX family
- Dialogue management queries



Architecture

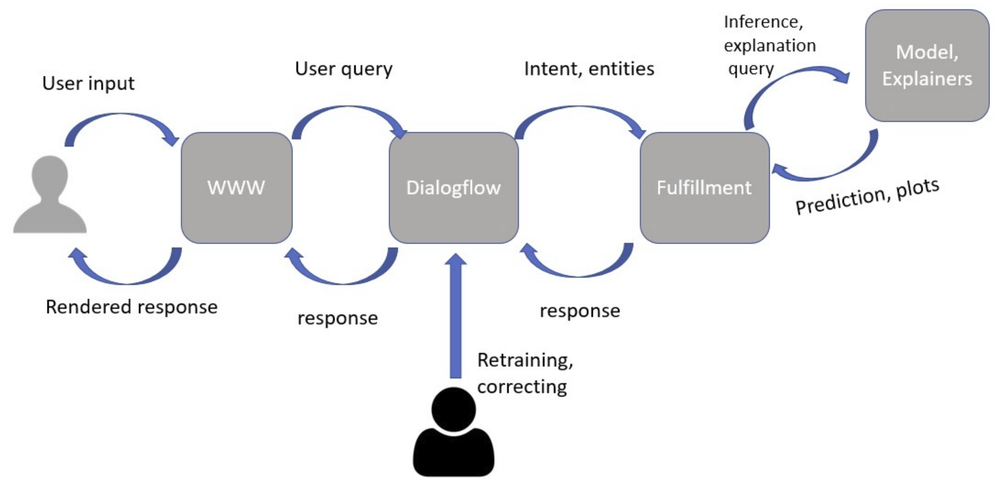


Figure 1: System architecture.

Results

Most of the explanatory questions fall into one of the following groups:

- "What-if" questions, such as "What if I'd been older?"
- General explanation queries, such as "Explain it to me" or "Why"
- Features relevance questions
- Maximizing prediction queries, e.g. "What should I do to increase my chances?"
- Local explanations queries, i.e. questions about the predictions for the similar passengers
- Contrastive explanations, here explanation for the difference in predictions for two passengers

Dataset

All conversations are logged to the Google Stackdriver. The anonymous dialogues along with the metadata are available on request.

Conclusions

- Conversational interaction with a model helps explore user concerns and questions about the Machine Learning blackboxes.
- There are several frequent patterns among the user queries and any successful explanation system should implement them. This helps to address the user concerns, aid the understanding and building the trust in the model as well as facilitate spotting any flaws of the deployed Machine Learning system.

References

[1] Przemyslaw Biecek. Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, 19:1–5, 2018.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, page 1135–1144. ACM Press, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL http://dl.acm.org/citation.cfm?doid=2939672.2939778.

[3] Kacper Sokol and Peter Flach. Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5785–5786. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/836. URL https://doi.org/10.24963/ijcai.2018/836.

Acknowledgements

This work was financially supported by NCN Opus grant 2016/21/B/ST6/0217.

- ✉ kuzba.michal@gmail.com
- https://kmichael08.github.io