Simplifying and Popularising Explainable and Interpretable AI

Explainable AI has always been a tricky subject within the field of computer science. While large effort has been put into developing faster and more effective algorithms for various forms of artificial intelligence such as machine learning, the results they produce are sometimes puzzling if interpreted manually.

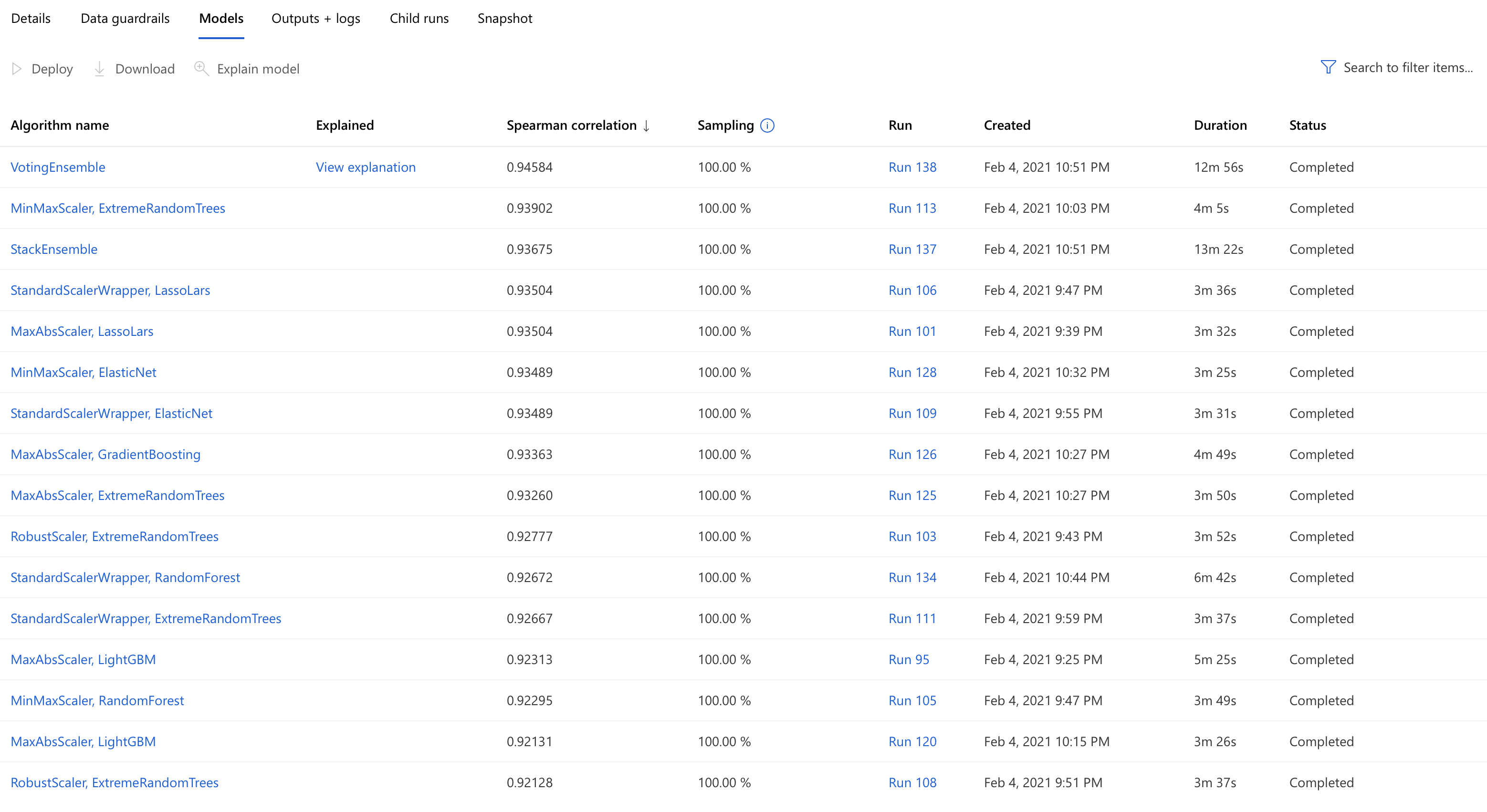
The VizAI project aims to develop a software application as a portable, intuitive and accessible solution to explaining one of the most popular forms of artificial intelligence: machine learning models and datasets. The concept of a specialised model explainer module is common, however most of them does not take usability into consideration, making the explainer module as puzzling to use as the machine learning models they are being used on. As such, it can be said that VizAI is a unique AI explainability project with both usability and feature robustness as goals for the explainer application it aims to develop.

In order to successfully achieve these aims, heavy research into projects related to AI explainability must be done. The VizAI team started by looking into explainability module and libraries in Python such as ELI5 and LIME. These modules and libraries are popular and well-maintained open source projects that see regular contribution, meaning their functions and reliability are indisputably solid.

Despite being relatively intuitive and easy to setup and use, most of these libraries still require a moderate understanding in the Python programming language and sometimes the machine learning models themselves, meaning they are in direct contradiction to one of the project goals: to offer a foolproof solution to interpreting AI behaviours that can be used even by the less technically advanced users.

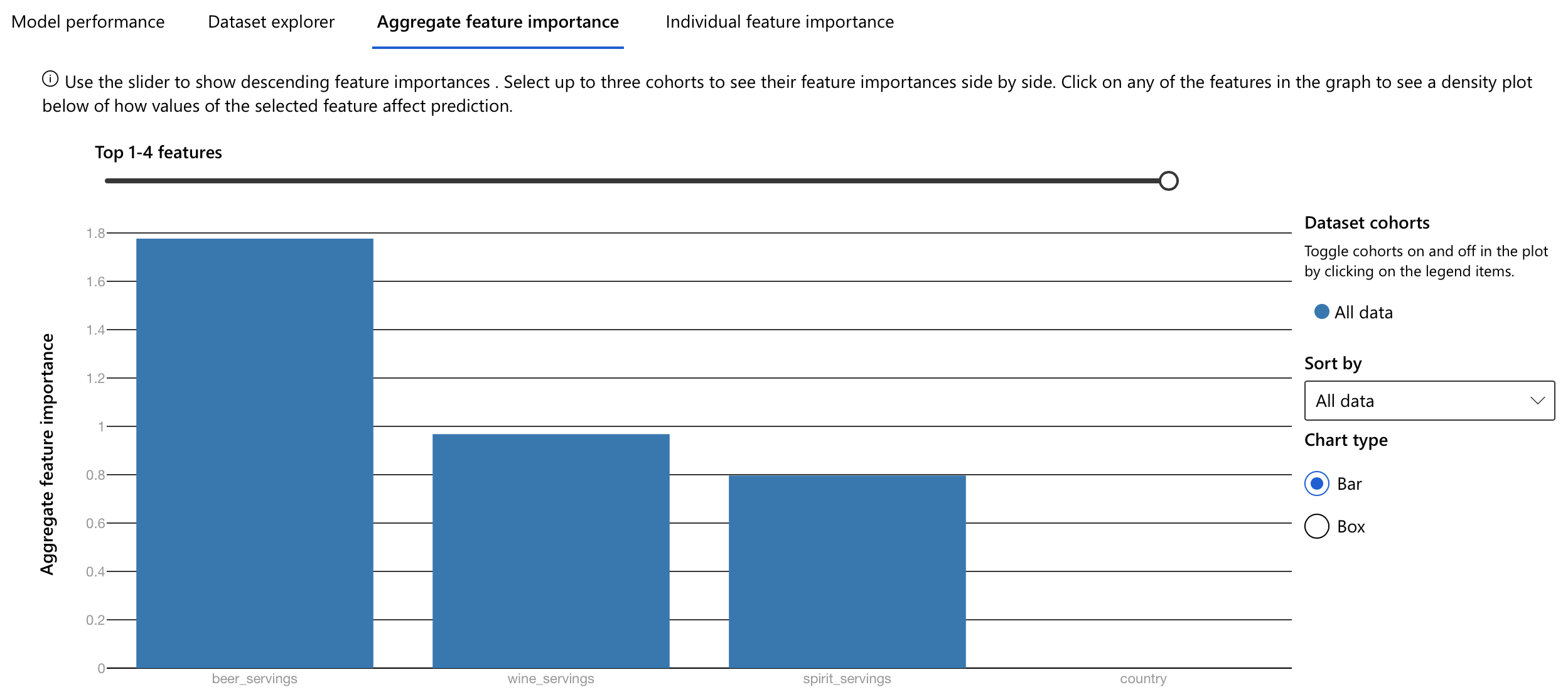
They provided good information nevertheless, as the team was able to learn how an explainability module functions on machine learning models and what kinds of explanation it would output. These knowledge helped the team tremendously in later stages of project development.

The team was able to review and benefit from reviewing one of Microsoft Azure Machine Learning Studio’s build-in function: the explainability module that is both a standalone tool and a submodule of the Automated ML functionality. Automated ML is Microsoft’s proprietary machine learning utility where by choosing the desired column to predict, the utility automatically trains models on a given dataset using different algorithms. The best model out of all attempts can then be explained using the explainability module.



A list of trained models using different algorithms

The team experimented with Automated ML and its explainability module to understand how to present the explanation in a clear and concise manner to the end users. The explainability modules supports data to be displayed in a variety of ways, the most helpful one being a bar chart displaying the aggregate feature importance. That is, which feature contribute the most to a model when making its predictions. As seen in the aggregate feature importance bar chart, when applied on a model to predict the total amount of alcohol consumed by a given country around the world, “beer\_servings” contribute the most to that prediction, meaning the model would base its prediction more on the amount of beer served in a country.



Aggregate feature importance bar chart, the highest bar on the chart is labelled “beer\_servings”

After carefully studying and analysing various types of AI explainability utilities and modules, the team revised and finalised its main project goals and officially started working on producing a prototype for a web-based application that would enable users to submit their own machine learning model and datasets and have them explained by clicking a single button.

It can be argued that with the advancement in the field of machine learning and its growing popularity, these tools and algorithms will eventually become widely distributed and become accessible to users that does not have specialised knowledge in machine learning. Therefore, it is necessary for them to be able to use a much simpler tool to explain the already complex machine learning models they might be using.

Unfortunately, the Python libraries mentioned above was deemed not simplistic enough to serve as the basis for the VizAI project.