SFU BIG DATA

CMPT 733 Big Data Science - Template for Capstone Project Idea

Explain Data and Interpretable Machine Learning

How are we doing

Description

This project is an opportunity for students at SFU to work with a product team at Tableau and may lead to future collaboration. We look forward to getting input from students about how we can improve our feature; the output from this project has the potential to impact future decisions.

Our team's first feature is Explain Data, which allows users to surface explanations with contextually relevant information about an aggregate data point. The statistical methods and interpretable machine learning that go into features like these are critical concepts every data scientist should be aware of. However, currently it still remains a challenge to take advantage of machine learning to share relevant contextual insights to those without ML knowledge, such that they can utilize and interpret the results. Explain Data is Tableau's first feature in this space.

"Explain Data gives you a new window into your data. Use it to inspect, uncover, and dig deeper into the marks in a viz as you build, explore, and analyze your data. When you select a mark while editing a view and run Explain Data, Tableau builds statistical models and proposes possible explanations for the selected mark, including potentially related data from the data source that isn't used in the current view.
As you build different views, use Explain Data as a jumping-off point to help you explore your data more deeply and ask better questions."

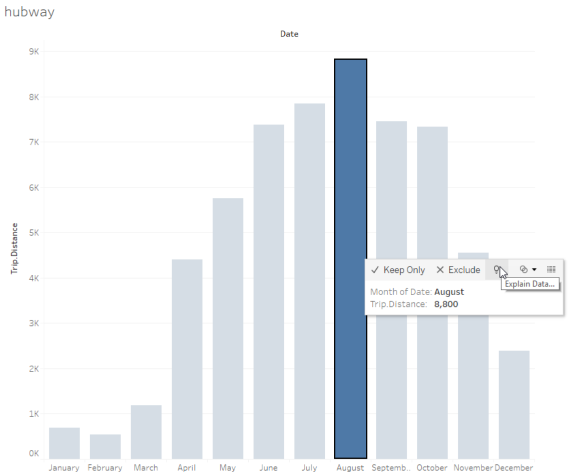
For more information please refer to our official documentation. I very much encourage students to watch

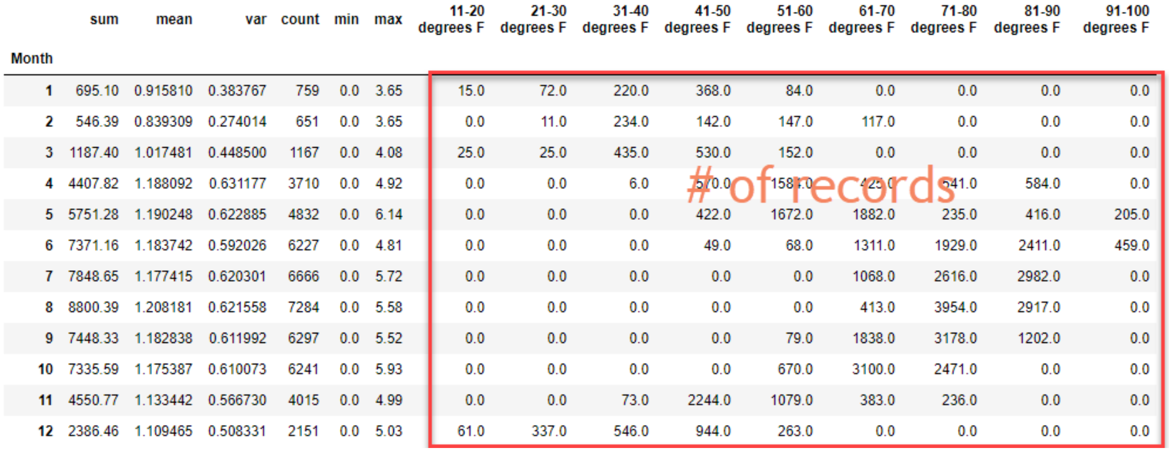
- The TC demo to get how powerful this feature could be
- Explain Data Internals: Automated Bayesian Modeling for an overview of different explainers and their implementation details.
- From Analyst to Statistician: Explain Data in Practice to understand the statement *"Use it to inspect, uncover, and dig deeper into the marks in a viz as you build, explore, and analyze your data".*

Of course, there are other tools for people with ML/DS knowledge. This tutorial on Kaggle gives an extensive review for some of the techniques to get started. **However please notice the different use cases from Explain Data and make adjustments accordingly**. Feel free to explore and experiment different ways to understand data, e.g. follow the workflow in the video linked above and interact with Explain Data to explore datasets (from e.g. Kaggle) of interest, as well as apply exploratory data analysis and modelling techniques to understand the data and find insights.

How do our explanations compare to what you've found on your own?? Compare the explanations you get from various approaches and find explanations that you deemed plausible and expected to have, but Explain data failed to uncover. What additional explanations does Explain Data have that you think are interesting and possibly enlightening? Or, on the contrary, boring or irrelevant? Feel free to also make a dashboard, story, or jupyter notebook to go over your process of surfacing and comparing explanations.

Let’s take an example where we can only pass in summary data for modelling. For example, given the viz in the interactive demo, where we want to explain why SUM(Trip.Distance) is high in August, when using Weather.Temp as a predictor, let’s say that we only have access to something similar to the following table. What other information can we pass that will be helpful for modelling? How would you model and approach the problem with aggregated data? Additionally, students are invited to consider the fact that Tableau is a general purpose data tool and therefore cannot make any assumptions about the dataset or its domain. More specifically, how can you generalize the model so that it can be applied to various datasets/visualizations?





	sum	mean	var	count	min	max	11-20 degrees F	21-30 degrees F	31-40 degrees F	41-50 degrees F	51-60 degrees F	61-70 degrees F	71-80 degrees F	81-90 degrees F	91-100 degrees F
Month															
1	695.10	0.915810	0.383767	759	0.0	3.65	15.0	72.0	220.0	368.0	84.0	0.0	0.0	0.0	0.0
2	546.39	0.839309	0.274014	651	0.0	3.65	0.0	11.0	234.0	142.0	147.0	117.0	0.0	0.0	0.0
3	1187.40	1.017481	0.448500	1167	0.0	4.08	25.0	25.0	435.0	530.0	152.0	0.0	0.0	0.0	0.0
4	4407.82	1.188092	0.631177	3710	0.0	4.92	0.0	0.0	6.0	570.0	1584.0	425.0	541.0	584.0	0.0
5	5751.28	1.190248	0.622885	4832	0.0	6.14	0.0	0.0	0.0	422.0	1672.0	1882.0	235.0	416.0	205.0
6	7371.16	1.183742	0.592026	6227	0.0	4.81	0.0	0.0	0.0	49.0	68.0	1311.0	1929.0	2411.0	459.0
7	7848.65	1.177415	0.620301	6666	0.0	5.72	0.0	0.0	0.0	0.0	0.0	1068.0	2616.0	2982.0	0.0
8	8800.39	1.208181	0.621558	7284	0.0	5.58	0.0	0.0	0.0	0.0	0.0	413.0	3954.0	2917.0	0.0
9	7448.33	1.182838	0.611992	6297	0.0	5.52	0.0	0.0	0.0	0.0	79.0	1838.0	3178.0	1202.0	0.0
10	7335.59	1.175387	0.610073	6241	0.0	5.93	0.0	0.0	0.0	0.0	670.0	3100.0	2471.0	0.0	0.0
11	4550.77	1.133442	0.566730	4015	0.0	4.99	0.0	0.0	73.0	2244.0	1079.0	383.0	236.0	0.0	0.0
12	2386.46	1.109465	0.508331	2151	0.0	5.03	61.0	337.0	546.0	944.0	263.0	0.0	0.0	0.0	0.0

# of records

After exploring the feature, we would be interested to hear:

- In addition to the explanations described in the above linked video, what other types of explanations would you want to surface?
- What other ways would you display explanations so that you or other users could understand it better?

Datasets

Please refer to The Best Public Datasets for Machine Learning and Data Science

Contact person

This feature is under active development, and we're always creating improvements and looking for feedback. Please reach out to kani@tableau.com if you encounter any problems or need clarification. Depending on the amount of communication, we'll likely host monthly office hours or invite students to the Tableau office to discuss more. We look forward to working with you!

Contributor of the Project Idea

Katrina Ni
Software Engineer – Machine Learning
Tableau Inc.