IDS707 Final Project

Data Story Title - Prediction Churn Reduction
Supplement (Slide Deck): Proofpoint Data Viz Final Project : https://www.canva.com/design/DAFUiDrdEFc/H5W13jfuhWefPvgt4b-ClA/view?utm_content=DAFUiDrdEFc&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton
Team Name: Team Proofpoint
Team Members: Dauren Bizhanov, Himangshu Raj, Satvik Kishore, Tigran Harutyunyan

A 1-paragraph overview of what your Capstone project is about

Our capstone client is Proofpoint Inc., an American cybersecurity company. Their services include a comprehensive spectrum of security, including email protection, online security, and private data security. Within this domain, they use machine learning for several tasks, like detection of fraudulent and risky emails. During the use of machine learning in this domain, our client has been facing the issue of prediction churn, i.e. inconsistency in model behavior upon retraining on new data.

Our project's primary goal is to develop a novel method that uniquely addresses our client's challenges and helps reduce their prediction churn. To achieve this goal, we divide it into the following tasks: we shall first compare three well-known methodologies with an integrated approach. Second, we shall develop a Python package that generalizes and implements these methodologies and is able to work regardless of underlying models. This will guarantee that our clients may utilize the program in their setting.

A statement describing what specific problem or question your visuals are meant to address

The visuals we have developed are meant to cover the goal of comparing the existing methods. To do this, we first dive into the concept of prediction churn and explain it through a hypothetical example. It then illustrates the mechanisms behind two of the churn reduction methods. We test the two methods on a dataset and the final section of the visualization covers the results from our experiments across four different metrics, to see which visual method is the best when it comes to reducing prediction churn.

A statement describing the intended audience for the visual data story

The intended audience for this visual data story are machine learning practitioners. These practitioners are well familiar with the challenges and pitfalls of training machine learning models. They are comfortable with the vector math that is needed to understand churn and may use the material from the slide deck in their own projects.

A statement describing how your stakeholders are expected to use the visual data story or interactive visualization Any details about your analyses or calculations that we might need to evaluate the validity of your insights, conclusions, or recommendations, especially if the details are not included in the other submitted materials

We anticipate that stakeholders will pay close attention to our slide deck since, in our instance, it is not just the visualizations that are significant, but also the content. If the stakeholder pays attention from the beginning of the presentation, they will have a thorough grasp of the methods that we implemented to address the issue.

A statement describing how your stakeholders are expected to use the visual data story or interactive visualization Anything else you think we need to know to evaluate your submission effectively

The client has not made their dataset and exact problem available to us yet. This visual data story serves to illustrate an example of a dataset and an attempt at reducing churn on it. This example helps the client understand the typical behavior and possible challenges that arise from using prediction churn reduction methods.