Feature Factory: a cloud-based platform for collaborative feature engineering

Micah Smith, Kalyan Veeramachaneni

ABSTRACT

Feature engineering is a critical step in a typical data science pipeline, in which raw variables are transformed to create features ready for inclusion in a machine learning model, and can be one of the most challenging aspects of a data science task. Feature Factory is a cloud-based platform for collaborative feature engineering, that allows independent data scientists to collaborate on a feature engineering task by viewing and discussing others' features in real-time, all within the familiar Jupyter Notebook interactive computing environment. Our software architecture includes a sophisticated auto machine learning backend that evaluates features registered by users in real-time. This backend abstracts away many aspects of a typical data science pipeline, allowing users to focus on feature engineering while still providing immediate feedback. We also use a tightly-integrated Discourse-based forum which facilitates discussion and collaboration among users. To facilitate genuine collaborative behavior, we introduce a compensation scheme that rewards users for high-quality features, hand-in-hand with participation in ideation and feedback discussions and accurate, well-written documentation. Preliminary results from user testing and experimentation show that data scientists increase their productivity using this collaborative platform, as opposed to working independently.