Group Discussion 1
BUDA 535
Team 5
Smith, Erwin, Frei, Patel & Strohl

The industry is awash in a wave of new tools and products designed to simplify some or all aspects of data science. These tools vary in terms of their usability, graphics and flexibility, but all are designed to improve the ability of organizations to collect, tidy, model and understand their data. The question is however, is this the end of data science and the rise of the machines?

Some of these products come from familiar names like Microsoft, IBM and Amazon. All of whom have established reputations for industrial scale computing and cloud storage. These companies leverage these advantages to create automated machine learning platforms that seem to be focused on training models with “one-click” and providing a simplified output score for the best model. All three emphasize the ability for wide-spread use across the organization to democratize data analytics by allowing subject matter experts to deal with data directly without knowing coding or statistics.

Less widely known products such as H20 focus a great deal on the tidying of data. They emphasize that 80% of a data scientist’s time is spent cleaning or munging the data to make it useful. H20 features data cleansing along with automated feature engineering to simplify that task. DataRobot also touts automatic feature engineering to recognize the specific techniques required by different algorithms. Their goal seems to also be to simplify the work of data analysts by automating the modeling and reporting.

While these packages differ in ease of use and scalability, they all focus on one thing. Simplifying the job of the analyst and trying to put more power into the hands of the subject matter experts. The purpose of autonomous machine learning is to free data scientists from repetitive tasks (like cleaning data) so they may spend more time on the analysis of the models.

Some may view these products as a means to eliminate or reduce humans in the process. After all, unsupervised learning techniques aim to discover patterns where no test data is available. Maybe that means we can use less expensive resources to crunch the numbers and generate the reports. But there are a few problems that autonomous machine learning hasn’t solved.

- Unsupervised Learning.  Unsupervised learning aims to discover patterns from unknown data.  Since the data is not known, the idea of a successful outcome may also be unknown.  The subjectivity of the outcome is where the expert's skills come to play.
- Reinforcement Learning.  With this, software learns through trial and error and receiving feedback from its actions.  This requires a definition of success and the subjectivity of an expert to accomplish.
- Feature Engineering.  It's been said that feature engineering is more art than science.  Feature engineering requires an expert to unearth the meaningful aspects of the process attempted to be modeled.  This requires imagination, creativity and most importantly, domain experience.

But none of these products actually claim to eliminate the role of humans in the process.  Instead, all of these products emphasize how they will free up the data scientists and subject matter experts to more fully engage with the outputs.  Some of the bigger products may hide their models behind scores, but others like H20 and DataRobot emphasize the elimination of non-value added tasks.  All of which are designed to provide the human with more time to analyze; gain insight from the variables; and most importantly to adapt their models and processes to explore new areas of the data.