EDA & Plotly Reflection

By: Benjamin Garnham, Jason Shiou, Lincoln Holt, Michael Maloney, Andrew Chang-Gu

Introduction

EDA, or exploratory data analytics, is an important type of data analytics that can provide many useful insights to an organization. The process of EDA involves first data exploration, then cleaning and transformation, and finally data visualization.

Exploratory Analytics

Exploratory data analytics is the process of investigating a dataset to discover patterns, and anomalies (outliers), and form hypotheses based on the understanding of the dataset. Within the exploratory phase there is a general learning cycle that follows curiosity, generating a hypothesis, transforming the data, visualizing the data, modeling to answer, and repeating. One big thing we learned in class is that the two phases of generating a hypothesis/transforming the data and visualizing the data are extremely important in delivering a sound analysis. The technical aspect of the data transformation is important as it is the source of truth. The visual aspect is key in delivery as it allows for the data to be understood by less technical folk, ultimately leading to mass interpretation.

In the corporate world each day exploratory analysis is becoming more democratized. What used to be the function of the business intelligence "BI" department is now spread throughout different areas of the companies. The main skill required in companies for exploratory analysis is the command of SQL, with over 80% of business analysts, a non technical position, reporting that they use it (Stack Overflow Survey, 2020). Another popular tool that has taken over are BI tools like Tableau and Power BI. These tools provide the possibility to non-technical people to plot and explore big databases as they are used to with Power Point plots and Excel pivot tables and their market is expected to grow over 10% per year for the next 5 years (Verified Market Research, 2021). This democratization of exploratory analysis is allowing companies to exploit the data they have collected and to better direct their future data science efforts.

Data Cleaning and Transformation

Data cleaning and transformation are almost always needed to make data capable of being analyzed by machine learning algorithms. Raw datasets often include anomalous entries, missing data points, incorrect formats, and other issues. The method chosen to clean the dataset for machine learning should be informed by the goal of the data analysis. For example, if we intend to perform anomaly detection on a dataset, we should not remove anomalous data entries. To account for missing data points, our class examined the use of multiple methods such as backfill, forward fill, and linear interpolation, where we take data points from adjacent rows to fill in missing data. These methods should only be used if adjacent data points are expected to be similar to the missing data point, if not, an average across the column could be used or the data point can be discarded. Data transformation includes the creation of categorical variables from initial inputs: for example, transforming the category variable of an item type to multiple variables that contain the true-false values of each category, ie. "is_cat_1" "is_cat_2" to allow regression analysis.

In the early days of data cleaning and transformation, manual entries were corrected using programs like Microsoft Excel – what a pain! Today, the *pandas* library in Python is commonly used to explore, evaluate, and enhance data clarity. Further, open-source projects, as well as SaaS-modeled companies, allow you to transform data at scale. Some of these include OpenRefine (a former Google product), Trifacta Wrangler, and Drake. These products take you through the cleaning process outlined below:

- Step 1: Remove duplicate or irrelevant observations
- Step 2: Fix structural errors
- Step 3: Filter unwanted outliers
- Step 4: Handle missing data
- Step 5: Validate and QA

Data Visualization

Data visualization is an important part of exploratory analytics. By generating summary statistics for numerical data in the dataset and creating various graphical representations (i.e., charts, tables, visuals, etc.), it allows presenters to better understand and explain the data even to non-technical audiences. One of the most helpful tools to do this within python is the plotly library. This open-source package has many different options of how to represent all types of data, including geographic, time-based, or even 3+ dimensional data.

For people who are interested in learning more about all the different ways of leveraging plotly, a good source is the library documentation (https://plotly.com/python/). Another good source that people can leverage is the quick guide on best practices on how to effectively leverage different chart types (https://www.columnfivemedia.com/resources/data-visualization-101-how-to-design-charts-and-graphs/). This is just as helpful as the library documentation as sometimes you need to think about how to best present the visualization as well in order to clearly convey your point.

Conclusion

The processes of exploratory analytics, cleaning and transformation, and data visualization are not only necessary prerequisites to the successful application of machine learning and AI algorithms on real world data, but can also provide valuable intelligence to the organization by itself. The tools and techniques for these actions are constantly improving and becoming accessible, putting data analytics capability within reach of more and more organizations.