Assignment 3.1: Impacting the Business with a Distributed Data Science Pipeline (Parts 1-2)

Authors: Leonid Shpaner, Jose Luis Estrada, Kiran Singh
Company Name: Street Infrastructure Solutions (SIS)
Company Industry: Streets Infrastructure/Transportation
Company Size: 3 (startup)

Abstract

The city of San Diego has become reliant upon a streets Overall Condition Index (OCI) that was designed and implemented by the United States Army Corps of Engineers. The company will provide recommendations to implement cost savings solutions.

Problem Statement

The city of San Diego has decided to "spend $700,000 to survey the condition of every street in the city so repairs and upgrades can be geared toward increasing social equity, fixing many long-neglected roads and boosting opportunities for bicycling" (Garrick, 2021). The challenge is to identify viable targets (streets) for future infrastructure projects for the city of San Diego. A high caliber consulting service that our company provides is instrumental for handling the following task. Classification of streets in above par conditions is a crucial step in establishing project feasibility. The city's future depends on it.

Goals

1. Predictive Analytics: Predict street viability presence/likelihood (good/fair vs. poor)
2. Prescriptive Analytics: Identify cost effective solution to expand infrastructure projects
3. Informative: Inform City of San Diego of the outcome in a timely manner (by 5/18/2022)

Non-Goals

While we will endeavor to provide recommendations and viable solutions that hinge on sound and proper data analytics, it is not in our capacity to “fix” issues including but not limited to traffic, parking meters, or real-estate assets or valuation.

Data Sources

Data will be stored on AWS service S3 Bucket that will communicate with AWS Sagemaker. The three files will be uploaded to S3 bucket.

- Streets Overall Condition Index (OCI): csv file 30,712 entries and 12 columns
  https://data.sandiego.gov/datasets/streets-overall-condition-index/

- Street Repair Projects: csv file with 23,433 entries and 19 columns
  https://data.sandiego.gov/datasets/streets-repair-projects/

- Traffic Volumes: csv file with 12,432 entries and 10 columns
  https://data.sandiego.gov/datasets/traffic-volumes/

Data Exploration

- Data will be stored in an S3 Bucket that will communicate with AWS Sagemaker visa vie AWS Athena to create the database and leverage SQL queries to combine the three files.
- During the exploration phase (EDA), column names, datatypes, missing values, and size/shape of the dataset will be initially documented. Bias exploration will determine the extent and/or effect

of imbalance data by looking at the target feature of “oci_desc” which provides information on the street quality with good, fair, and poor conditions, respectively. This effect will be measured both numerically and represented visually on a bar graph. Moreover, histograms of all of the numeric features on the joined dataframe will be produced to establish the presence of degenerate distributions. One boxplot examining streets’ overall condition index (OCI) will be presented. Lastly, a correlation table will follow suit to examine potential sources of multicollinearity.

- SQL by way of Athena (through PyAthena) will be used to ingest the data. Pandas will be used to read in the sql queries visa vie the `pd.read_sql()` function. More broadly, the Pandas library will be used to read-in and explore the dataframe(s), while matplotlib and seaborn will be used for visual exploration (graphs). Moreover, an additional helper tool for table visualization (prettytable) will be used too.
- GitHub Repository (notebook): https://github.com/lshpaner/sd_streets/blob/main/sd_streets.ipynb
- GitHub Repository (main): https://github.com/lshpaner/sd_streets

Measuring Impact

- Specifically within the target variable, it is expected that the “good,” “fair,” and “poor” street condition classes will be cast to dummy variables. It is additionally expected/entirely possible that the three variables will be narrowed down to two whereby a binary classification will follow (i.e., good condition vs. poor condition (0,1)).
- Provided that certain machine learning methods/models have the ability to extract predicted probabilities, this will allow for a new column with such metrics to be feature engineered at the culmination of predictive modeling.

Security Checklist, Privacy and Other Risks

- No PHI, PII, user behavior, nor credit card data will be stored or processed since the information presented/provided herein is a matter of public record.
- This application will read/write to the following public s3 bucket: s3://waterteam1/raw_files/
- Bias by way of class imbalance will be considered/addressed in order to assuage the potential effects of overfitting some or all of the machine learning methods/models that will be explored. Re-balancing the classes where an imbalance exists by oversampling or undersampling is one method of addressing this roadblock.
- One ethical concern that should be addressed is overfitting/underfitting the data commensurate with the initial notions of the viability/efficacy of the dataset at large.

References

Garrick, D. (2021, September 12). San Diego to spend $700K assessing street conditions to spend repair money wisely. *The San Diego Union-Tribune.* https://www.sandiegouniontribune.com/news/politics/story/2021-09-12/san-diego-to-spend-700k-assessing-street-conditions-to-spend-repair-money-wisely