Assignment 4.1: Impacting the Business with a Distributed Data Science Pipeline (Part 3)

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

An S3 bucket is created in which a parent folder directory `raw_files` has three separate folders for each respective .csv files. The data is stored in an S3 Bucket that will communicate with AWS

Sagemaker visa vie AWS Athena, a serverless “interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL” (Amazon Web Services, n.d.) to create the database and combine the three files into one single dataframe `df.`

Exploratory Data Analysis (EDA)

During the exploration phase (EDA), column names, data types, missing values, and size/shape of the dataset are initially documented in a new cell block. There are initially a total of 28 columns (features) and 23,005 rows that are a combination of floating point numbers, objects, and integers. Information on whether or not each respective column contains any null or missing values is represented herein. At this stage, missing values are uncovered in date_moratorium (4,426), date_start (1), date_end (7), street_name (16,874), and total_count (16,874), respectively.

Summary Statistics and Outlier Detection

Table 1 shows the summary statistics of the target variable, overall street condition, re-indexed by range values.

Table 1

Overall Condition Index (OCI) Summary

Count	Mean	Std	Min	25%	50%	75%	Max
23005	74.79141	16.78405	0	66.3	79.06	87.3	100

Note. The mean is lower than the median, suggesting a negatively skewed distribution on the target variable.

Whereas the low (Q1 - 1.5*(IQR)) and high (Q3 + 1.5(IQR)) outliers are found to be 34.8 and 118.8, respectively, omitting these does not benefit long-term project goals. Resistance versus sensitivity to outliers in this endeavor is part and parcel of further analysis.

Data Ingestion

SQL by way of Athena (via PyAthena) is used to ingest the data and Pandas is used to read in the sql queries visa vie the `pd.read_sql()` function. More broadly, the Pandas library is used to read-in and explore the dataframe(s), while matplotlib and seaborn are used for visual explorations (graphs). An additional helper tool for table visualization (prettytable) is leveraged for added visual appeal.

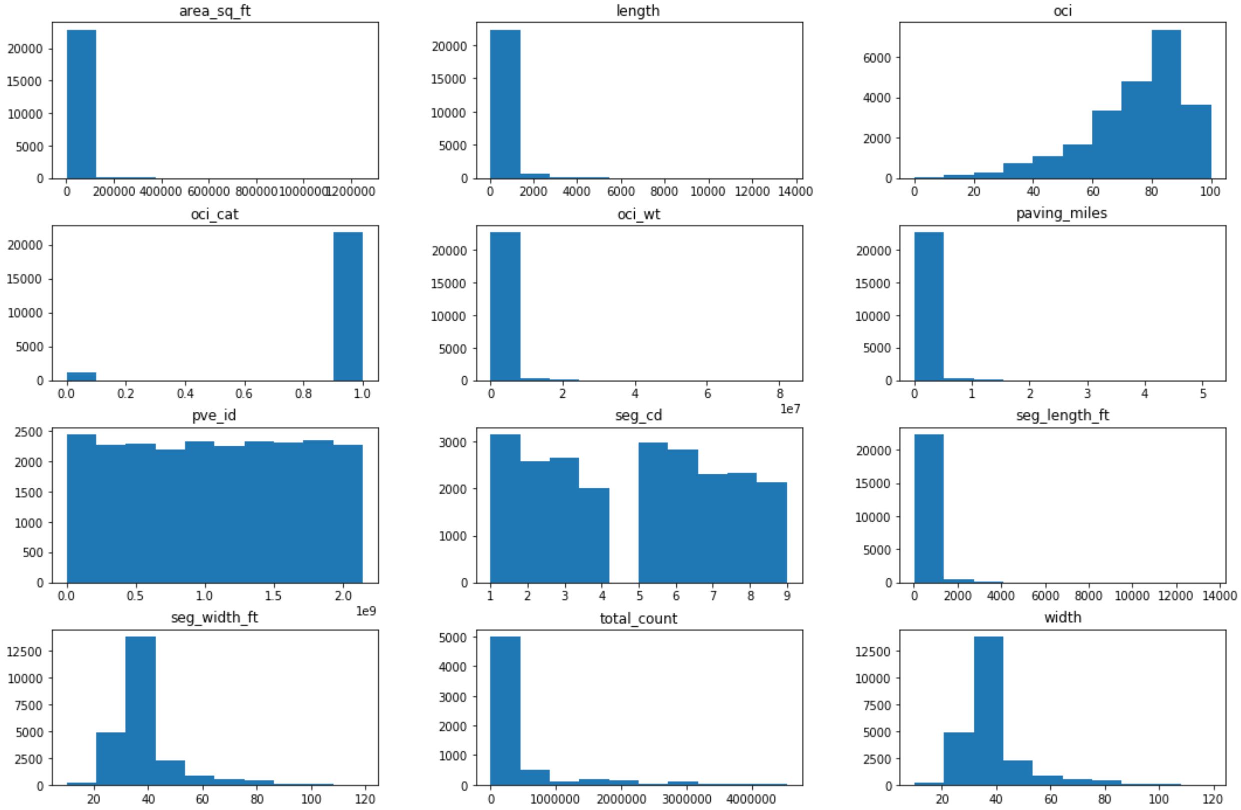
GitHub Repository Information

Main notebook (.ipynb file): https://github.com/lshpaner/sd_streets/blob/main/sd_streets.ipynb
Main notebook (.pdf file): https://github.com/lshpaner/sd_streets/blob/main/sd_streets.pdf
Master Link to main repository: https://github.com/lshpaner/sd_streets

Moreover, histograms of all of the numeric features on the joined dataframe are produced to establish or detect the presence of degenerate distributions. One accompanying boxplot examining streets' overall condition index (OCI) is presented visually, but illustrates the same behavior (summary statistics) that has already been depicted in Table 1. Figure 1 shows distributions of all of the numerical features from the entire dataset.

Figure 1

Histogram Distributions



Note. Area in square feet, length, paving miles, and segment length in feet all exhibit right-skewed distributions. The OCI categorical feature is negatively skewed where there is a class imbalance between the 0 and 1 classes, respectively. This is supported by the ensuing Bias Exploration section.

Notwithstanding, all accompanying proportional measurements (i.e., height, width, length, etc.) are true and proper records acquired by the city of San Diego. No suitable transformation (normalization or standardization) is required in order to avoid the potential adverse effect of a high bias, low variance model whereby " a higher bias would not match the data set closely" (Wickramasinghe, 2021).

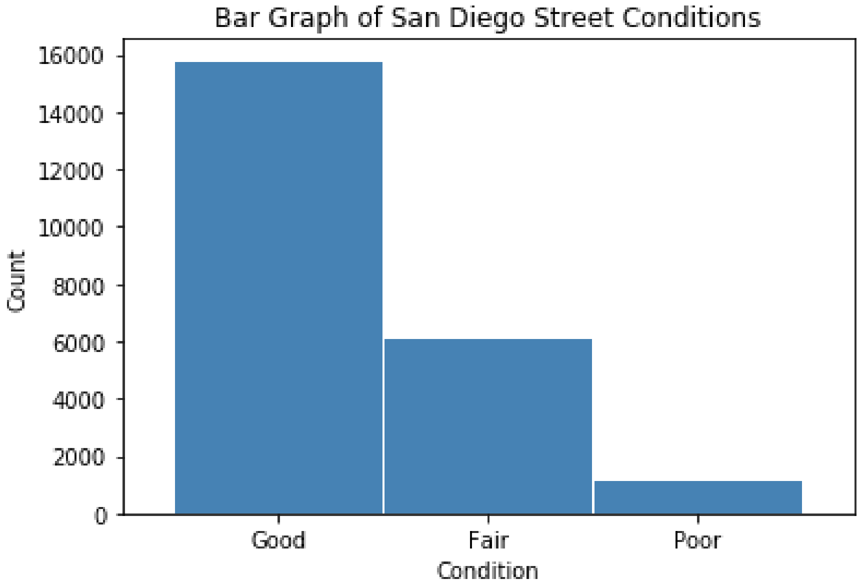
Pavement identification and total count are of no value and should thus be removed from the dataset.

Bias Exploration

Bias exploration helps 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 is measured both numerically and represented visually on a bar graph. There are a total of 23,005 streets of which 6,105 streets are in fair condition, 15,758 streets in good condition, and 1,142 streets in poor condition. Figure 2 shows this categorical distribution on an accompanying bar graph.

Figure 2

Bar Graph of San Diego Street Conditions



Whereas a method can be used to classify street conditions into multiple classes, it is easier to re-classify streets in “fair” and “good” condition into one category in comparison with the poor class. This, in turn, becomes a binary classification problem. Thus, there are now 21,863 streets in good condition and 1,142 in poor condition (only 5% of all streets). This presents a definitive example of class imbalance.

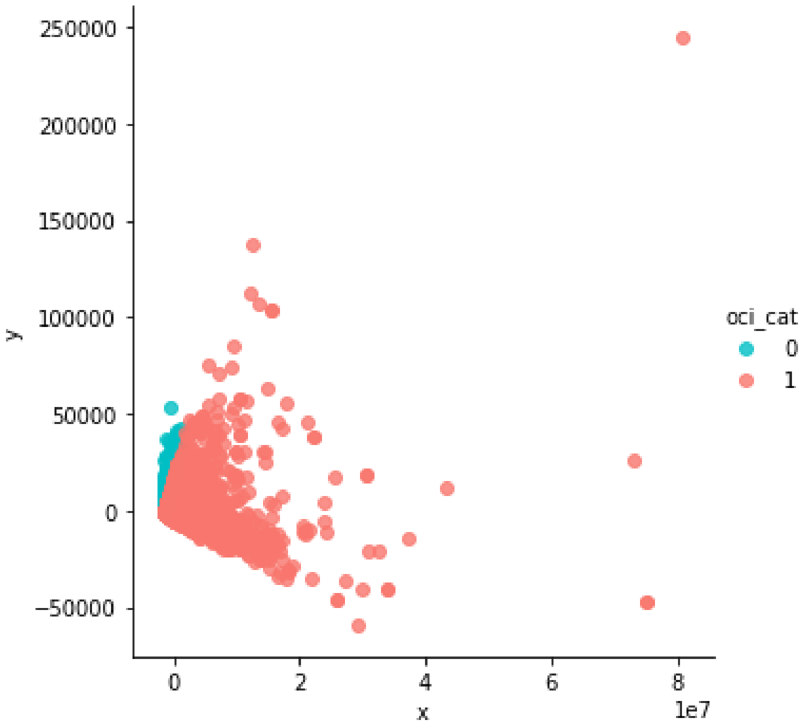
Class Imbalance

Multiple methods for balancing a dataset exist like “undersampling the majority classes” (Fregly & Barth, 2021, p. 178). To account for the large gap (95%) of mis-classed data on the “poor” condition class, “oversampling the minority class up to the majority class” (p. 179) is commenced. However, such endeavor cannot proceed in good faith without the unsupervised dimensionality reduction technique of Principal Component Analysis (PCA), which is carried out “to compact the dataset and eliminate irrelevant features” (Naseriparsa & Kashani, 2014, p. 33).

In this case, a new dataframe is reduced down into the first two principal components since the largest percent variance explained exists therein and because these principal components are depicted on the ensuing two-dimensional (x,y) scatterplot in Figure 3. A two-dimensional analysis is the most parsimonious one for illustrating additional visual confirmation of a class imbalance by coloring the classes in two distinct colors (light blue and pinkish red).

Figure 3

Class Imbalance in Streets’ Overall Condition



One final endeavor in exploratory data analysis yields a triangular correlation matrix, an important step for examining the relationship between predictor variables and determining multicollinearity based on a threshold of a pearson correlation coefficient *r* = 0.75. Based on this criteria, area in square feet, oci_wt, length, width, and paving miles are columns that are earmarked for subsequent removal, but this may not be necessary if too many features are to be removed. This is further discussed in the pre-processing section.

Measuring Impact

Specifically within the target variable, it is expected that the “good,” “fair,” and “poor” street condition classes being cast to dummy variables may slightly over-generalize street conditions by placing more emphasis on poorer conditions. To this end, these three variables are narrowed down to two whereby a binary classification follows suit (i.e., good condition vs. poor condition (0,1)).

Provided that certain machine learning methods and 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 is considered and 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.

Data Preparation and Data Scrubbing visa vie Pre-Processing

Date_start and date_end are subsequently removed after being concatenated into one uniform feature (date_days). Total count and street name represent the same information, and are unimportant features that are dropped from the dataframe altogether. Moreover, any duplicate columns are removed from the dataframe at large; this is an extra measure for avoiding post-join feature redundancy.

Predictive models only work with numerical values; therefore, categorical features such as func_class (function class), pvm_class (pavement class), and status are transformed into numerical values by mapping dictionaries of categorical values in ascending order. Creation of dummy variables is supported by the following information. For example, in the function class feature, the residential, collector, major, prime, local, and alley functional classes are converted to categorical values (1-6). Similarly, in the pavement class feature, AC Improved, PCC Jointed Concrete, AC Unimproved, and UnSurfaced pavement classes are converted to values ranging from 1-4.

Lastly, the current status of the job (i.e., post construction, design, bid/award, construction, and planning) is converted to categorical values between 1-5. Features with no additional value are removed. Columns with explicit titles (i.e., names) and non-convertible/non-meaningful strings are dropped. Redundant columns (columns that have been cast to dummy variables) are dropped in conjunction with the index column which holds no value in this work.

For context and clarification columns with identifying information that are dropped include the project id, pavement event id, segment id, and project title. Additionally, the project manager’s email and phone number are removed to protect sensitive information in accordance with strict compliance standards and because this information cannot be ingested into a viable machine learning algorithm.

Balancing the Dataset

The adaptive synthetic sampling approach is leveraged in conjunction with the Synthetic Minority Over-sampling (SMOTE) technique to “balance the class distribution and increase the variety of sample domain” (Naseriparsa & Kashani, 2014, p. 33). This allows for the minority class to be more closely matched (re-sampled) to the majority class for an approximately even 50/50 weight distribution. This results in a larger dataset. Whereas previously there were 23,005 rows, there are now 43,702.

Train, Test, Validation Splits

To avoid overfitting, the main dataset is split into three respective component parts (dataframes), which will work to train, test, and validate the final model (Solawetz, 2020). Using sci-kit learn, the split is divided by 70%, 15%, and 15%, respectively. Whereas specifying a stratify parameter allows the split function “to choose any data in the given dataset, causing the splits to become unbalanced” (Fregly & Barth, 2021, p. 181), a random_state is set to 777 to ensure reproducible results.

References

Amazon Web Services. (n.d.). *Amazon Athena.* https://aws.amazon.com/athena/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

Fregly, C. & Barth, A. (2021). *Data Science on AWS.* O'Reilly.

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

Naseriparsa, M. & Kashani, M.M.R. (2014). Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset. *International Journal of Computer Applications, 77*(3) 33-38. https://doi.org/10.5120/13376-0987

Solawetz, J. Train, Validation, Test Split for Machine Learning. *Roboflow.* https://blog.roboflow.com/train-test-split/

Wickramasinghe, S. (2021, July 16). Bias & Variance in Machine Learning: Concepts & Tutorials. *Bmc blogs.* https://www.bmc.com/blogs/bias-variance-machine-learning/