Sutenee Nopachinda
Udacity Machine Learning Nanodegree – Capstone Project Proposal
July 12, 2020

Crash Attribute Segmentation of New Jersey State Vehicle Accident Data

The following proposal outlines a project of using K-Means clustering algorithms to identify crash patterns across New Jersey counties by crash attribute segmentation.

Domain Background

Every year, the New Jersey Department of Transportation (NJDOT) publishes vehicle accident data which is derived from New Jersey Police Crash Investigation Report forms (NJTR-1) submitted by police officers. This data is used to support a variety of safety programs and planning efforts such as analysis of crashes at major intersections and calculations of fatalities and injuries in counties or municipalities.

In my work environment, the crash data is typically analyzed in a county or municipality scale using standard data analysis methods to calculate crash counts or injury counts. With these data analysis methodologies, it is difficult to examine crash patterns across the entire New Jersey State and its 21 counties. After completing the machine learning case studies in this program, I learned several algorithms that would help me explore this data in a macro level.

My motivation for this project is to explore data that I analyze daily with the machine learning algorithms I learned in this course. I want to learn more about attribute segmentation on a deeper level and see if I can utilize these new skills in a real-world data setting. I also want to improve on my data visualization skills and learn different tools to portray meaningful data to the end user.

Problem Statement

The goal of this project is to identify and explore crash patterns across counties in New Jersey. As there are over 20 columns of textual data in NJTR-1 forms and over five million data entries across 21 counties, it is difficult to use standard data analysis methods to compare crash data between multiple counties and identify correlations between multiple crash attributes. Typically in my work environment, this data is filtered in a linear workflow – first by a specific county or location, then applying crash attributes such as alcohol involvement, pedestrian involvement, or person severity rating. After learning different machine learning algorithms from the case study examples, I thought this project would be a great opportunity to explore the data using unsupervised learning algorithms that I learned.

Using the Population Segmentation case study as inspiration, Principle Component Analysis (PCA) methods and K-Means clustering algorithms will be used to identify major principle components between crash attributes across New Jersey counties. After extracting meaningful results from the training model's attributes, visualization tools will be used to portray the results that would be easy to understand for the end user.

Datasets

The datasets to be used in this project are the 2015-2018 Crash data tables publicly provided by NJDOT. The Crash table is a summary of a crash accident and includes data columns including location the accident took place, if alcohol was involved, the highest victim severity rating, and road conditions. Many values under these columns are numerical codes which are transcribed from New Jersey Police Crash Investigation Report forms.

As the goal of this project is to identify crash patterns across New Jersey counties, Crash tables for numerous counties will be used. For an initial test, up to five counties will be used for PCA and training a K-Means algorithm. Once the results and workflow are verified, the number of counties will be scaled up and the model will be re-trained again to identify crash patterns across all counties.

Solution Statement

To find crash patterns among different counties and municipalities, PCA methods and K-Means algorithms will be used for crash attribute segmentation. PCA and K-Means algorithms from Sagemaker, scikit, and PyTorch will be explored and their resultant clusters will be evaluated. Visualization methods including heatmaps and feature-level makeups will be used to compare results between the algorithms.

Benchmark Model

The benchmark models for this project will consist of models from Sagemaker, scikit, and Pytorch. PCA methods from each library will be used to compare dimension reduction results. After deciding on which transformed dataset to use, K-Means algorithms from each library will be used to explore how each algorithm clusters the transformed data.

Evaluation Metrics

This project will explore the results from Sagemaker, scikit, and PyTorch PCA methods and K-Means algorithms. In the dimension reduction process, the top principle components each algorithm chooses can be compared and the explained variance of each algorithm can be calculated. To evaluate resultant clusters from each K-Means algorithm, visualization methods such as heatmaps, cluster graphs, and tables will be used to compare cluster labels. A summary of differences and similarities will be provided, as well as a discussion on crash patterns that were discovered.

Project Design

The following sections are the proposed workflow for this project.

Data Loading, Cleaning, and Preprocessing

The text-based data will be loaded Amazon S3 and a Juypter notebook will be used as an interface with the data. Columns that contain textual data such as "Description", "Case Number", and "Police Department" will be removed. Since machine learning models require numerical data and not categorical data, the data for each county and municipality will need to be preprocessed. The methodology for this will be to calculate percentages for every numerical code in each column, essentially flattening the data table.

For example, the column "Severity Rating" has five numerical codes that correspond to a person's injury condition. This column will be transformed into five columns which will contain the percentage of the corresponding code for the entire municipality.

County	Municipality	Severity_05	Severity_04	Severity_03	Severity_02	Severity_01
Atlantic	Hamilton	2%	7%	13%	21%	57%
Atlantic	Longport	8%	11%	21%	17%	43%

Dimension Reduction using PCA

PCA models from Sagemaker, scikit, and PyTorch will be used to reduce the number of features within the preprocessed dataset.

Feature Engineering and Data Transformation

Resulting features and explained variance from each PCA method will be evaluated. A total variance of 80-90% will be the target for selecting the number of principle components.

Clustering Transformed Data with K-Means Algorithms

K-Means models from Sagemaker, scikit, and PyTorch will be used to cluster the transformed datasets. Using empirical data, a K value of 8 will be used initially. After extracting k-means model attributes, visualization tools such as heatmaps, tables, and component makeup histograms will be used to compare results between the three models. A summary of differences, similarities, and crash patterns will be provided. If time permits, I will deploy a data-driven web application which would utilize JavaScript visualization libraries to allow users to interact with my findings.