Summary

In this project, I am going to analyze the dataset of parking citations in Los Angeles by implementing various classifiers, such as logistics regression and decision tree, etc. Since the dataset contained a large amount of data, to obtain the most useful information, I first made a data processing to retain only the relevant columns and focus on only the brand of Toyota. Additionally, I separated the 'Issue Date' column into year and month column to make it more readable and rearranged to dataframe again.

In the main project, I utilized five different methods, Logistic Regression, Decision Tree, Naïve Bayesian, K-nearest neighbors, and K-means clustering, respectively. Overall, I would like to figure out the relationship between the specific body style of car and its fine amount. Therefore, I used the groupby function and sorted in a descending order to determine that the panel vehicle was my targeted variable.

In the first three methods (Logistic Regression, Decision Tree, and Naïve Bayesian), I set the data of the year 2017 as my training data and the year 2018 as my testing data. I also set the fine amount as the independent variable (x) and replace any NaN value with zero. For the dependent variable (y), I used a function to return one if the body style is a panel (PA) and zero otherwise. Moreover, I calculated the accuracy, confusion matrix, true positive rate, and true negative rate for each classifier, respectively. The results showed in the Appendix.

For K-nearest neighbors, I first determined the optimal value of k by assessing a list of k-values in a for loop***. The result indicated that the optimal k-value was 3 with the highest accuracy of 0.915108 among the list. Next, inputting the optimal k-value of 3 into the K-nearest neighbors classifier, I figured out its accuracy, confusion matrix, true positive rate, and true negative rate, respectively.

Lastly, for K-means clustering, I first filtered the dataset to contain the data for both the year 2017 and the year 2018. And then, I implemented a 'knee' chart to find out the best k-value (see Appendix). In the case, the best k-value was 2 because when the k-value was less than 2, the slope of changing in total within-clusters sum of squares (y value) was sharper. By contrast, when the k-value was greater than 2, the slope became flat and stable. Furthermore, using the best k-value, I separated the dataset into two clusters and found the percentage of body style in each cluster and determined whether there was a pure cluster in any cluster.

Instruction

Running the code, you simply click the run button. It may take a while since the dataset is large

***Due to a large amount of data, the program may run a while (approx. 30 minutes). For convenience, please see the Appendix below for the best k value, accuracy, confusion matrix, TPR and TNP

Appendix

Logistic Regression

The accuracy for year 2018 by implementing logistic regression is 0.915104

	Predicted: Other	Predicted: Panel (PA)
Actual: Other	TN = 0	FP = 28209
Actual: Panel (PA)	FN = 0	TP = 304069

True positive rate is 1.0

True negative rate is 0.0

Decision Tree

The accuracy for year 2018 by implementing decision tree is 0.915095

	Predicted: Other	Predicted: Panel (PA)
Actual: Other	TN = 156	FP = 28053
Actual: Panel (PA)	FN = 159	TP = 303910

True positive rate is 0.99947709

True negative rate is 0.00553015

Naïve Bayesian

The accuracy for year 2018 by implementing Naïve Bayesian is 0.909810

	Predicted: Other	Predicted: Panel (PA)
Actual: Other	TN = 198	FP = 28011
Actual: Panel (PA)	FN = 1957	TP = 302112

True positive rate is 0.99356396

True negative rate is 0.00701904

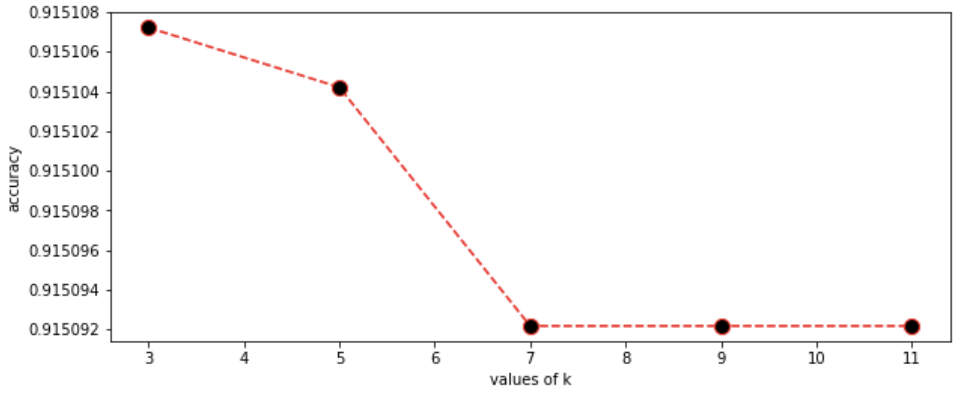
K-nearest neighbors

The accuracy for year 2018 by implementing K-nearest neighbors is 0.915107

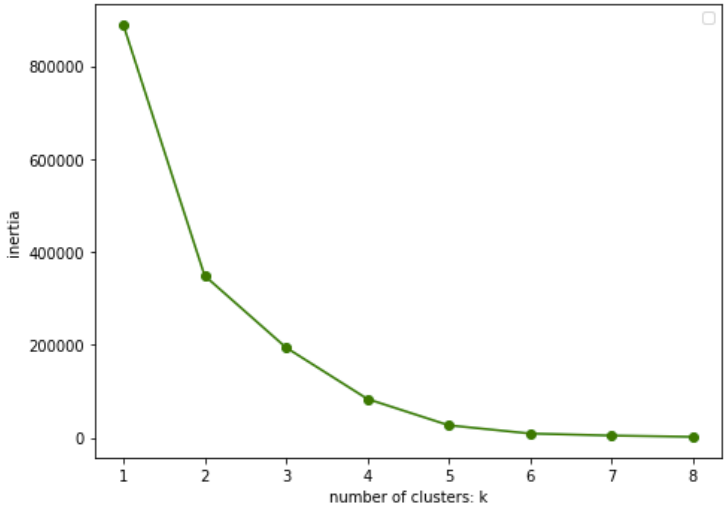
	Predicted: Other	Predicted: Panel (PA)
Actual: Other	TN = 1	FP = 28208
Actual: Panel (PA)	FN = 0	TP = 304069

True positive rate is 1.0

True negative rate is 3.545e-05



K-means clustering



In the first cluster, the percentage of PA body style is 0.91 and the percentage of other body style is 0.09
In the second cluster, the percentage of PA body style is 0.9 and the percentage of other body style is 0.1
My first clustering for PA body style is a pure cluster