Machine Learning Approaches to Predict Crime Rate in the City of Los Angeles

Supervisor: Youngmin Ha
Group member: Jieyi Yao, Rongsheng Zhang, Ziqing Chen

ABSTRACT

Intro:

– Machine learning (ML) algorithms is widely accepted by scholars as a prediction tool, a very dearth of research has attempted to measure the crime rate and identify contributing factors to perpetration of a crime. Filling this research gap, this paper models the crime rate of City of Los Angeles by employing four machine learning (ML) algorithms. This poster attempts to identify the best-performing ML algorithms by comparing the performance of the ML algorithms. Of the four ML algorithms tested, the CatBoost is selected as the best-performing algorithms.

Method:

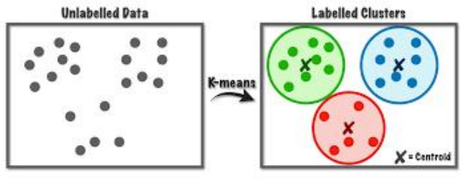
- Clustering
- Linear regression
- SVC (Support Vector Classifier)
- CatBoost

Aim:

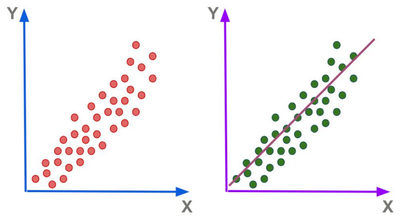
– Identify the best-performing ML algorithms by comparing the performance of the ML algorithms.

METHOD

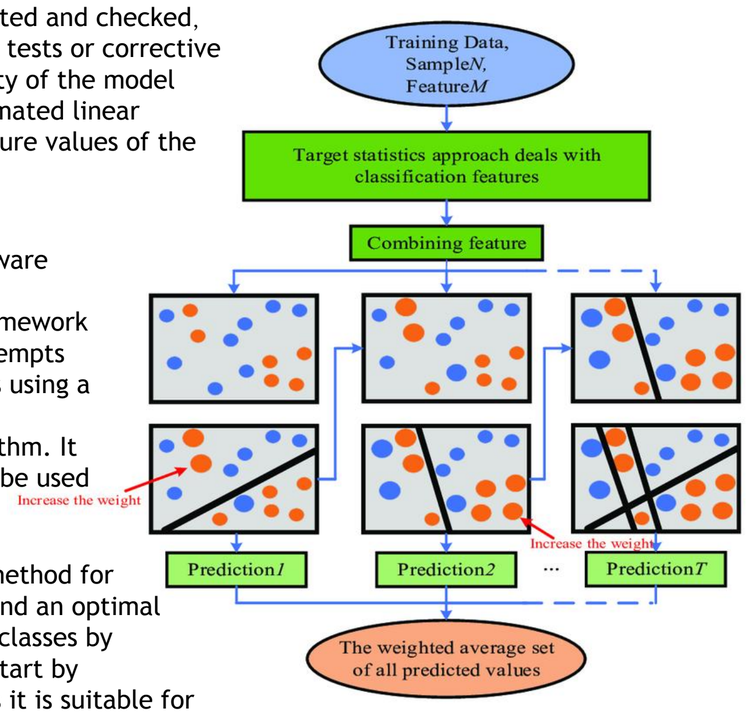
- Clustering:
  - is a widely used unsupervised learning technique that aims to group similar data points together based on their inherent similarities. Our approach is based on the popular K-means clustering algorithm. To determine the optimal number of clusters, we employ the elbow method. This method involves running the K-means algorithm for a range of values of K and calculating the sum of squared distances between the data points and their respective centroids.



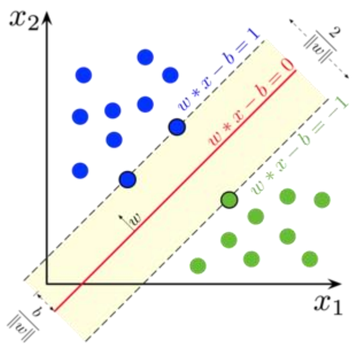
- Linear regression
  - is a widely used statistical technique that aims to model the relationship between a dependent variable and one or more independent variables. The general form of the model is given by:
    $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$
  - where Y represents the dependent variable,
  - $X_1, X_2, \ldots, X_n$ are the selected independent variables, $\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients, and ε represents the error term. Thirdly the Model will be evaluated and checked, we apply appropriate diagnostic tests or corrective measures to ensure the reliability of the model results. Finally, we use the estimated linear regression model to forecast future values of the dependent variable.



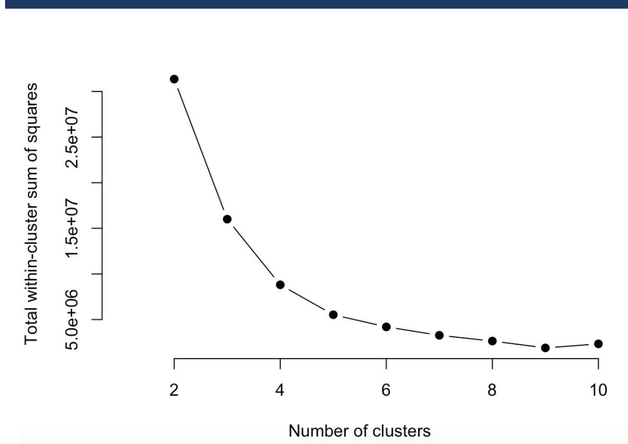
- CatBoost
  - CatBoost is an open-source software library developed by Yandex. It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm. It models built using catboost can be used for predictions.

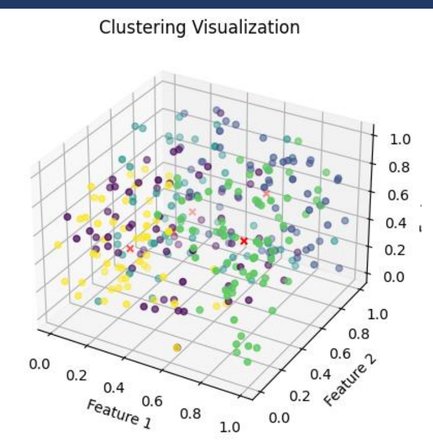


- SVC
  - is a popular supervised learning method for classification tasks that aims to find an optimal hyperplane to separate different classes by maximizing the margin. We also start by preprocessing the data to ensures it is suitable for training the SVC model. To determine the most relevant features, we employ techniques such as correlation analysis, feature importance ranking, or dimensionality reduction methods like Principal Component Analysis (PCA). We train the SVC model by using the training set. Then we evaluate the performance of the SVC model using appropriate metrics recall or F1 score. SVC provides insight into the importance of different data points through support vectors. We assess the generalization performance of the SVC model by comparing the predicted labels with the true labels of the test set.

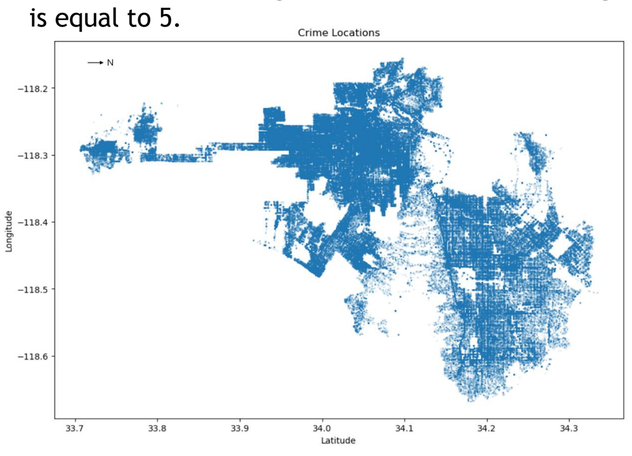


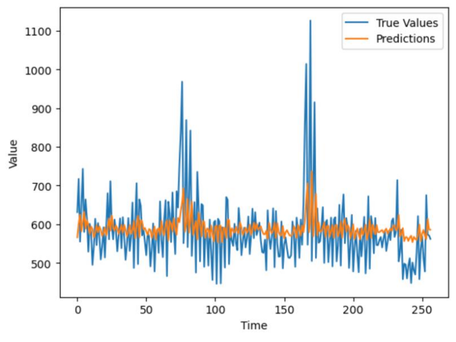
COMPARISON ALGORITH



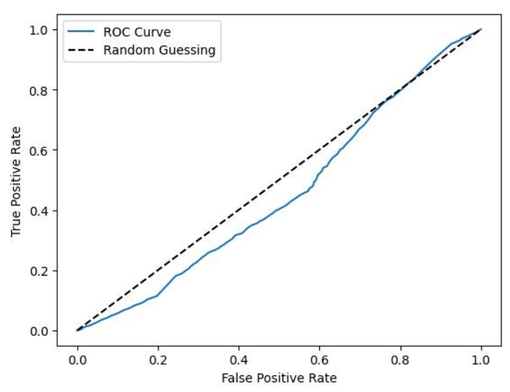


The elbow curve is plotted, which shows the relationship between the number of categories and the overall within-class sum of squares. The fviz_nbclust function is used to automatically determine the optimal number of classes and output the result. The figure above is the result, it was plotted by r language. The line to the left of 5 cluster falls steeply and the line to the right of 5 cluster tends to be gentle. So, we find that the suitable cluster is equal to 5.

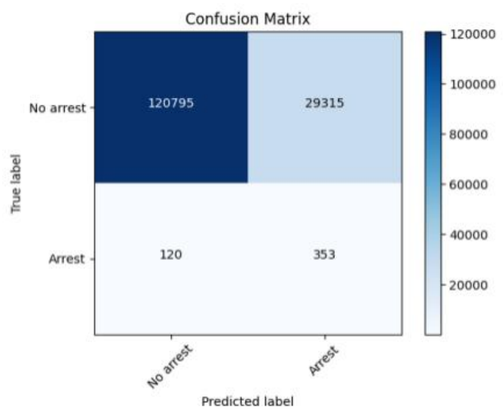


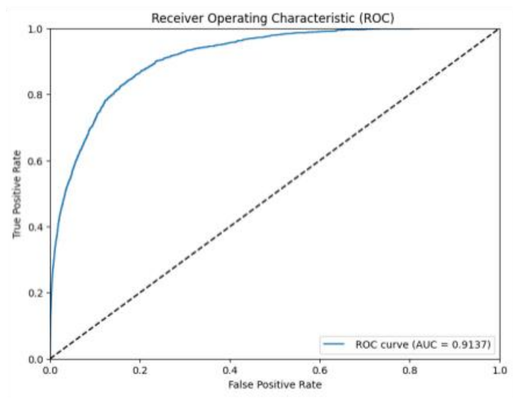


Mean Squared Error: 89.43665652302413



The prediction accuracy obtained by the SVC model was observed to be only 0.45353623155627787. This suggests that the model has limitations in accurately recognizing crime data, indicating the need for substantial improvements in its performance for crime data recognition





From the model evaluation section, this model achieved 91.37% accuracy on the test set, meaning that correctly predicted 91.37% of the samples.In the confusion matrix, it is evident that the model exhibits outstanding performance in identifying the negative class (class=0), correctly predicting a substantial number of 120,795 samples, with only 29,315 samples being misclassified as positive. However, on the positive class (class=1), the model demonstrates a comparatively lower accuracy. Although it correctly predicted 120 samples, there were instances where 353 samples were erroneously classified as negative.

CONCLUSTION

Comparing the accuracy of linear regression, logistic regression, SVC, and Cat Boost algorithms in predicting the data, it can be found that Cat Boost has an advantage in prediction accuracy and is the algorithmic model that is more in line with crime data prediction.

In this project, we use the knowledge of machine learning to classify, analyze and predict crime data. It provides data support for the police to find similar crimes when analyzing crime data, to reduce crime and enhance social stability.

There are still many deficiencies in this project, such as the accuracy of the data cannot be checked in the data collection, the experiment has a certain degree of randomness, and the results cannot predict the cause of the crime more carefully.

To dig deep into factors such as criminal motives, establish comprehensive and accurate early warning information. It is necessary to combine field investigations and theoretical research to further establish more effective crime prediction related models. Luckily, we also tested the ability of using machine learning methods to solve daily problems.

REFERENCES

1. Xiaochuan SUN, Tianliang LU, 2019. Application of Data Weighting Optimization Based on Clustering in Crime Prediction. JISUANJI YU XIANDAIHUA, (6), pp.55-59.
2. Hilpisch, Y., 2014. Python for Finance: Analyze big financial data. " O'Reilly Media, Inc.".
3. Mueller, J.P. and Massaron, L., 2021. Machine learning for dummies. John Wiley & Sons.
4. Lewinson, E., 2020. Python for Finance Cookbook: Over 50 recipes for applying modern Python libraries to financial data analysis. Packt Publishing Ltd.