Final Project Report
Data Science 602: Introduction to Data Analysis and Machine Learning

Predicting Crime Types Based on Historical Data: A Machine Learning Approach

Business Problem:

The dataset Crime Data from 2020 to the Present contains details regarding crimes in various cities around the City of Los Angeles between January 2020 and the present day. The dataset has attributes for the location, date, and time of the offense and the type of crime that was committed. Based on the location, victim sex, victim race, victim age, date, and time of the incident, a machine-learning model can be created using the dataset to forecast the sort of crime that will be committed.

The goal of this project is to create a machine-learning model that can identify the kind of crime being committed based on the location, the time and day it occurs, and other information. This can help law enforcement officials effectively utilize their resources and prevent crimes from happening in the future.

Project Approach:

CRISP-DM, short for Cross-Industry Standard Process for Data Mining, is a widely utilized methodology for data mining and machine learning projects. It provides a structured approach to guide data scientists and analysts through the various stages of a data mining project, from understanding the business problem to deploying the final model.

The CRISP-DM methodology consists of six main phases:

Business Understanding: This initial phase focuses on comprehending the project's objectives and requirements. The emphasis is on understanding the business goals, identifying key stakeholders, and defining the problem that the data mining project seeks to address.

Data Understanding: In this phase, the available data is collected, explored, and analyzed to gain insights and familiarity with its characteristics. Data quality issues, inconsistencies, and missing values are assessed. This step helps determine the data's suitability for the project and identifies any necessary data preprocessing steps.

Data Preparation: This phase involves preparing the data for modeling. It includes data cleaning, feature selection, feature engineering, and data transformation tasks. The goal is to ensure that the data is in the appropriate format and structure for modeling.

Modeling: In this phase, various machine learning models are selected and built using the prepared data. The models are trained and evaluated using appropriate evaluation techniques. The focus is on developing models that best address the business problem and meet the project's objectives.

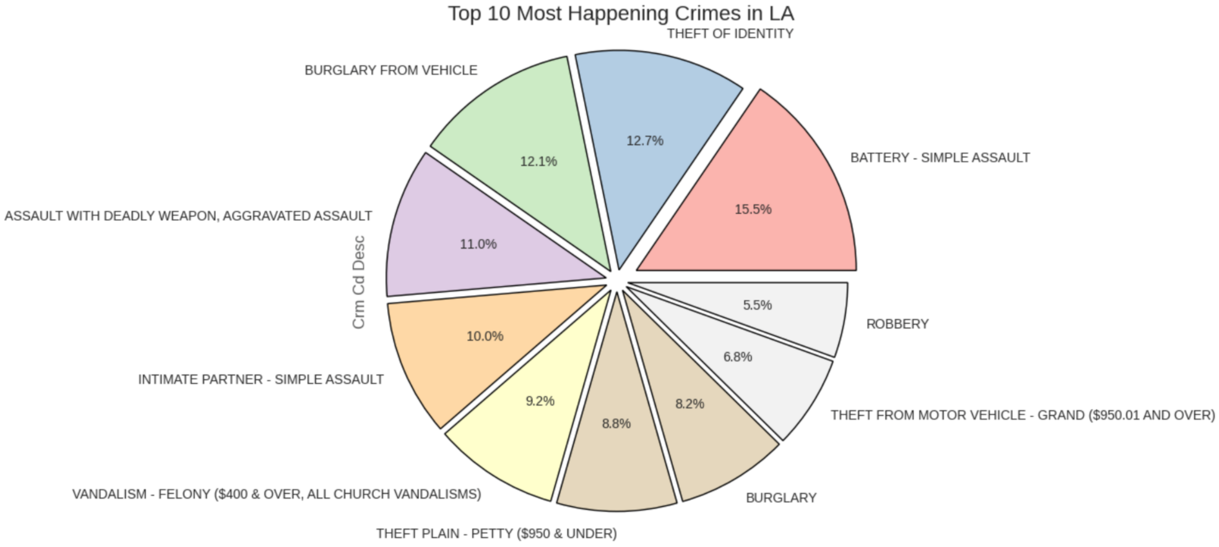
Evaluation: The models built in the previous phase are evaluated and compared based on predefined evaluation metrics and criteria. The goal is to assess the models' performance, determine their strengths and weaknesses, and select the best-performing model(s) for further refinement.

Deployment: In the final phase, the selected model(s) are deployed into a production environment, integrated with existing systems, and made available for end-users or stakeholders. This phase also includes creating documentation, providing user training, and establishing a monitoring system to track the model's performance in real-world scenarios.

The CRISP-DM methodology is iterative, allowing for revisiting previous phases as new insights or requirements emerge. It provides a systematic and flexible framework to guide data mining projects, ensuring a well-structured and organized approach from start to finish.

Type of Modeling: Classification

Due to the large dataset size, I have identified the top three crimes that happen in LA and predicted them using Random Forest Classification and Extreme Gradient Boosting.



The top three crimes are:

Battery- Simple Assault

Theft of identity

Burglary From Vehicle

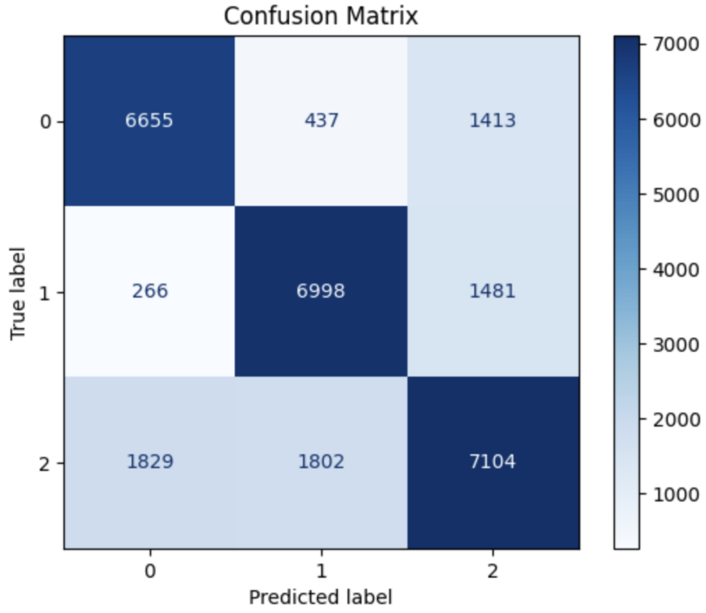
Target variable - Crm Cd

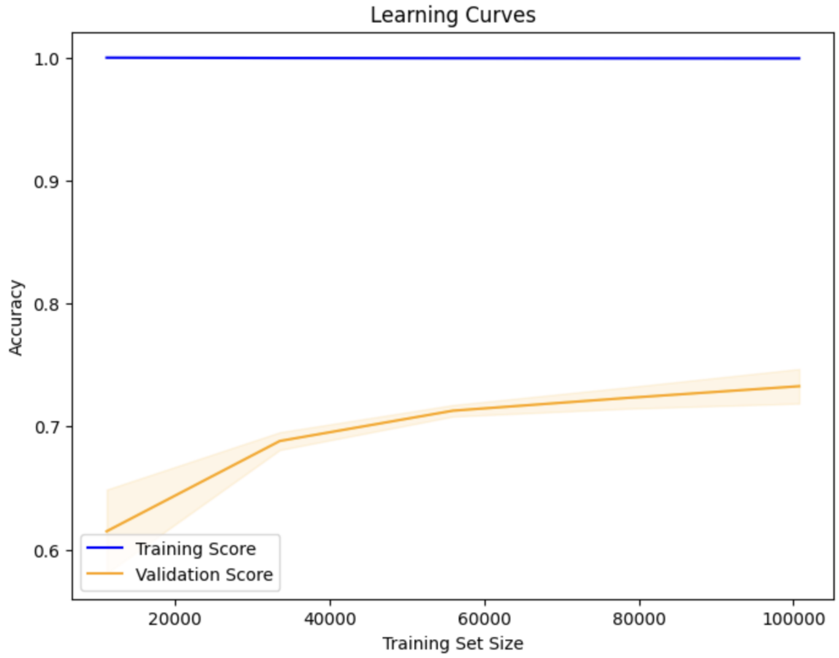
Feature Variable: Area, Vict_Age, Vict_Sex, Vict_descent, Premis cd, Time_occ, Day_of_week, Premis Cd

Results:

Results of Random Forest Classification:

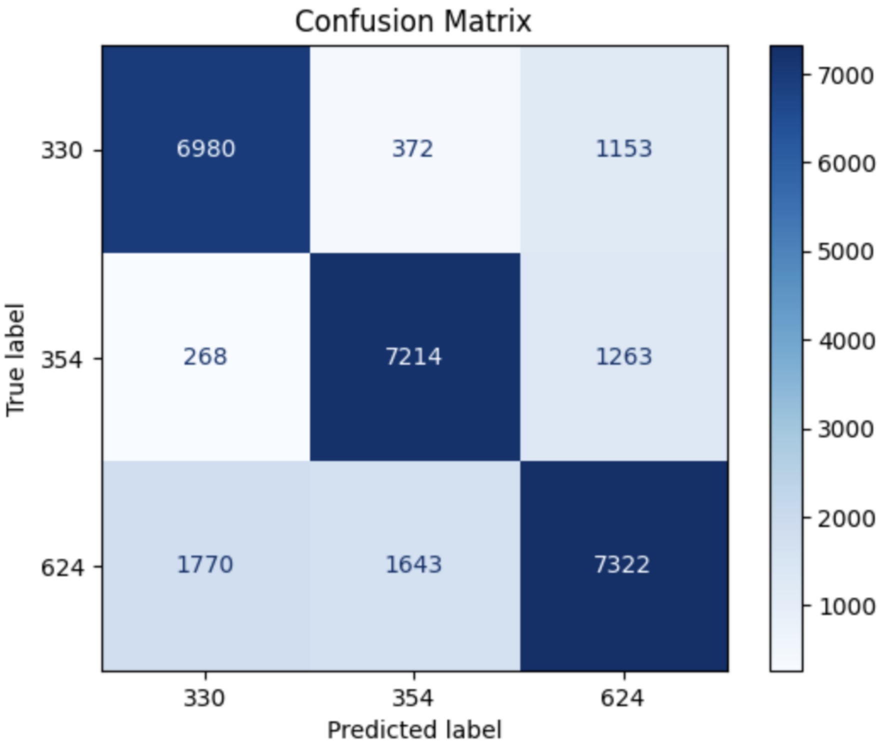
Accuracy: 0.7417187779167411

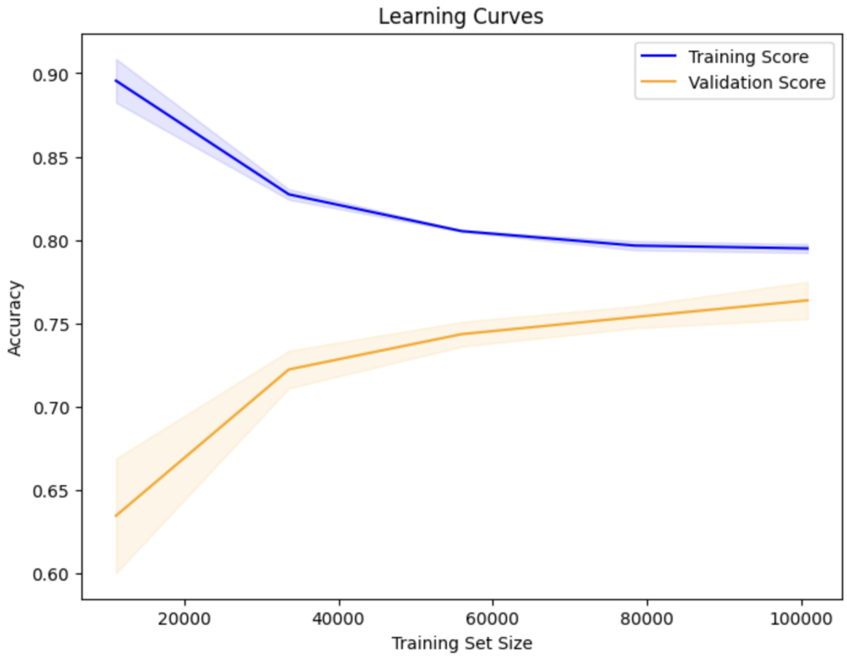




Results from XGboost:

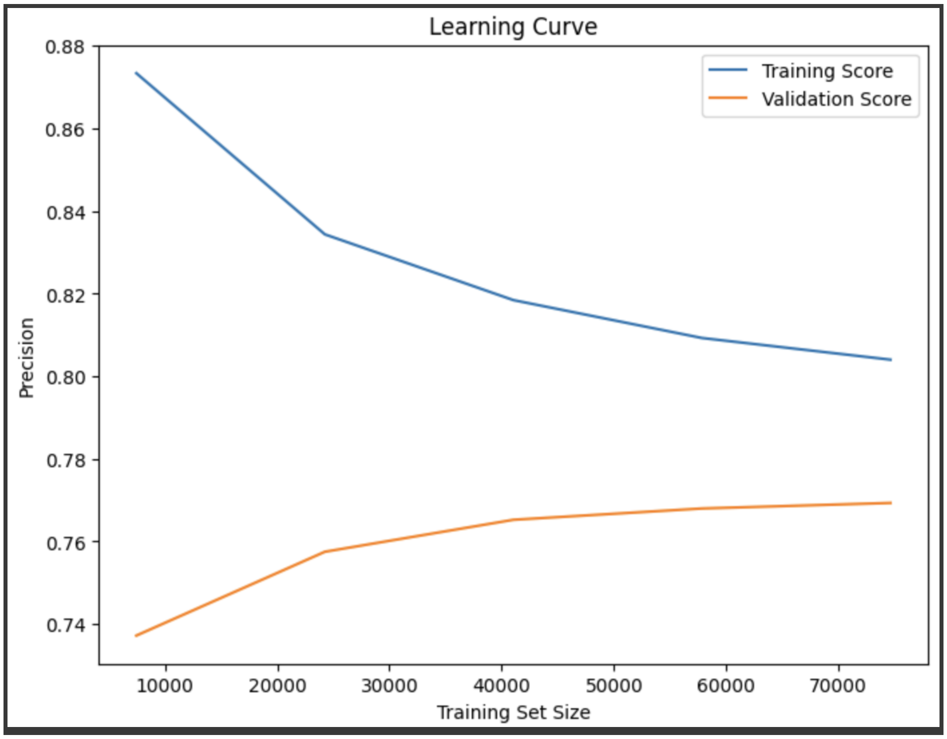
Accuracy: 0.7688404502412006





Stacking (Decision Tree classifier, Random Forest classifier, and XGBoost):

Precision: 0.7701786972292215



Model Performance: The Extreme Gradient Boosting (XGBoost) algorithm demonstrated the highest accuracy among the tested models, achieving an accuracy of 0.76. This indicates that XGBoost was able to effectively capture the complex relationships between the input features and the crime types, resulting in more accurate predictions compared to other models.

Random Forest Classification: The Random Forest Classification algorithm also performed well, achieving an accuracy of 0.74. While it showed a slightly lower accuracy compared to XGBoost, it still proved to be a reliable model for crime type prediction. Random Forest models excel in handling high-dimensional data and capturing feature interactions, making them suitable for this task.

Model Stacking: The ensemble technique of stacking, which combined the predictions of a Decision Tree classifier, Random Forest classifier, and XGBoost, resulted in further improvement, achieving an accuracy of 0.77. By leveraging the strengths of multiple models, stacking enhanced the predictive power and performance, surpassing the individual models' accuracy.

Conclusion:

Based on the results, it can be concluded that the machine learning approach using XGBoost, Random Forest Classification, or a stacked ensemble can effectively identify crime types based on location, time, and other relevant factors. The predictive accuracy of 0.77 achieved by the stacked model indicates its potential for practical implementation in assisting law enforcement officials in resource allocation and crime prevention strategies.