Crime Prediction of category in San Francisco

National Taipei University of Techonology
Department of Industrial Engineering and management
黃鈺凱, *109370210*
陳國誌, *109370211*

Introduction

San Francisco, once home to the notorious Alcatraz prison from 1934 to 1963, now thrives as a hub of technological innovation. Despite its transformation into a tech mecca, the city continues to grapple with significant social challenges. Rising wealth inequality, severe housing shortages, and the ubiquitous presence of costly digital gadgets contribute to an ongoing tapestry of urban crime.

In this analysis, we explore a dataset encompassing nearly 12 years of crime reports across San Francisco's diverse neighborhoods—from the serene streets of Sunset to the bustling corners of SOMA, and from the affluent Marina to the vibrant Excelsior. The task is to predict the category of crime based on temporal and spatial data.

Problem statement

- **Data Wrangling**: This stage involves the data quality and taking necessary steps to cleanse the dataset, ensuring it's free of inconsistencies or errors.
- **Data Exploration**: We delve into the dataset to understand its variables and develop insights, which will help inform subsequent analyses.
- **Feature Engineering**: In this phase, we create additional variables from the existing ones to enhance the dataset's predictive power or interpretability.
- **Data Transformation**: This step prepares the dataset for machine learning algorithms by standardizing or transforming the data as needed.
- **Training/Testing Data Split**: We partition the data into training and testing sets to evaluate our models' performance and fine-tune their parameters.
- **Model Selection**: The ultimate goal is to develop a predictive model that estimates the likelihood of each crime type based on location and date, allowing for effective resource allocation and preventive measures.

Data Overview

The training set includes the following columns:

- **Dates**: Timestamp of the crime occurrence.
- **Category**: Type of crime (the target variable for prediction).
- **Descript**: Detailed description of the crime.
- **DayOfWeek**: Day of the week when the crime occurred.
- **PdDistrict**: Police district where the crime occurred.
- **Resolution**: Outcome of the crime (e.g., arrest made, case closed).
- **Address**: Location of the crime.
- **X**: Longitude coordinate of the crime location.
- **Y**: Latitude coordinate of the crime location.

The training set consists of 878,049 entries, with each entry providing detailed information about a specific crime.
The test set includes the following columns:

- **Id**: Unique identifier for each record.
- **Dates**: Timestamp of the crime occurrence.
- **DayOfWeek**: Day of the week when the crime occurred.
- **PdDistrict**: Police district where the crime occurred.
- **Address**: Location of the crime.
- **X**: Longitude coordinate of the crime location.
- **Y**: Latitude coordinate of the crime location.

The test set also contains 878,049 entries, similar to the training set but without the **Category**, **Descript**, and **Resolution** columns.

Exploratory Data Analysis (EDA)

To facilitate analysis and modeling, we need to convert categorical variables into numerical values. Given the high number of unique values in some columns (e.g., Category with 39 unique values), using one-hot encoding is impractical. Instead, we use count encoding for categorical variables and mapping for ordinal variables.

We define a function to apply count encoding to specified columns. Count encoding replaces each category with its frequency count in the dataset.

Through this initial data exploration and preprocessing, we have transformed categorical features into numerical values, setting the stage for detailed exploratory data analysis (EDA) and subsequent predictive modeling. This foundational work allows us to effectively analyze and model the patterns within the crime data, leading to better insights and accurate predictions of crime categories.

Hypotheses

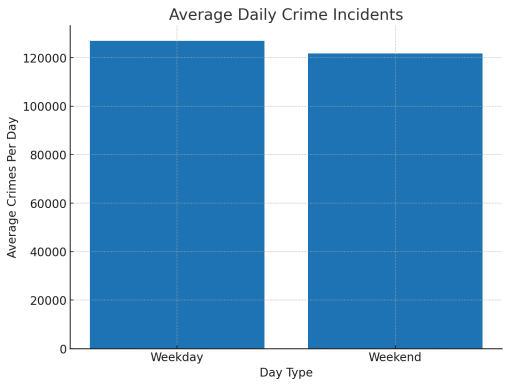
- **Hypothesis 1**: Crimes are more likely to occur on weekdays than weekends.
- **Hypothesis 2**: Crimes are more likely to occur during late night/early morning hours compared to other times of the day.
- **Hypothesis 3**: Crimes of similar types are more likely to occur in close proximity to each other.

Hypothesis 1: Weekday vs. Weekend Crimes

The hypothesis suggests that weekdays have a higher incidence of crimes compared to weekends. This is based on the idea that during weekdays, people are more active and mobile, creating more opportunities for crime. In contrast, weekends are typically quieter, with people spending more time at home.

Result:

- Average **weekday** crimes per day: 126,906.40
- Average **weekend** crimes per day: 121,758.50

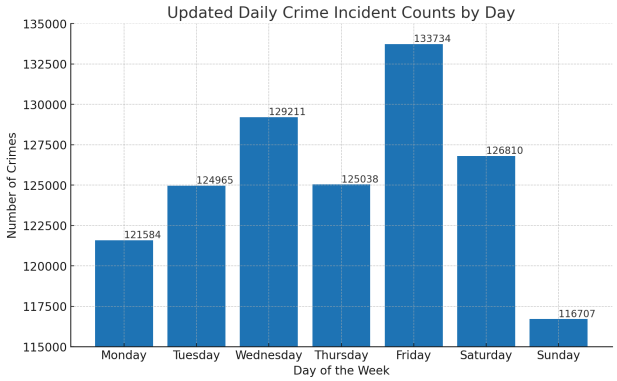


Findings

The analysis shows that the average number of crimes per day is higher on weekdays compared to weekends by approximately 4%. This supports the hypothesis that crimes are more frequent on weekdays.

Daily Crime Distribution

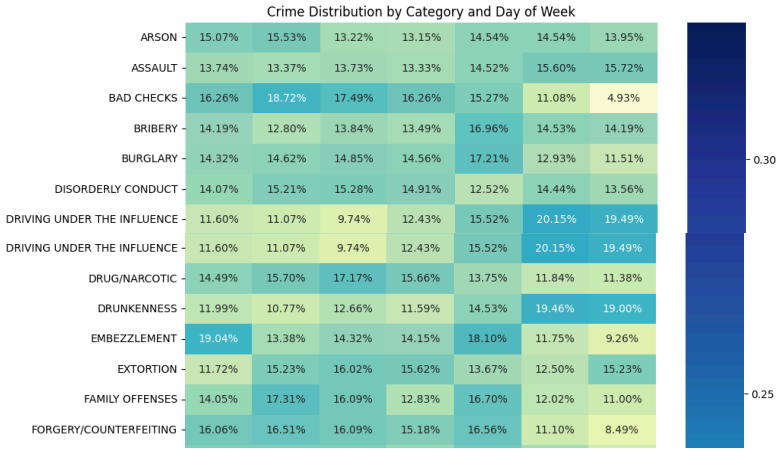
To further understand the distribution of crimes across the week, we examined the daily crime counts.



Pattern Analysis by Crime Category

We also explored whether each crime category follows the general pattern of higher weekday crime incidence.

By evaluating the crime categories, we identified whether they matched the overall pattern of higher weekday crimes.



Conclusion of Hypothesis 1

Out of 39 crime categories:

- **26 categories** show a higher frequency on weekdays.
- **13 categories** show a higher frequency on weekends.

This supports the hypothesis that crimes are generally more frequent on weekdays, although there are significant exceptions among specific types of crimes.

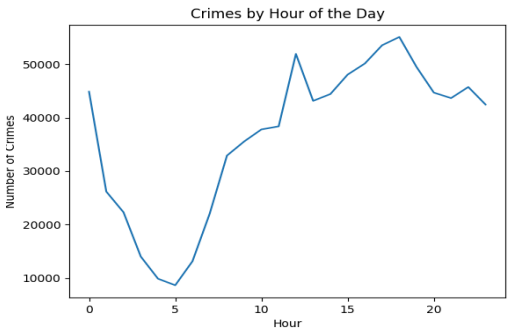
Hypothesis 2: Crimes and Time of Day

Late night and early morning hours are typically periods of reduced public activity and visibility, which might provide cover for criminal activities. The hypothesis posits that these times would see a higher incidence of crimes compared to other times of the day.

Data Analysis

To test this hypothesis, we divided each day into distinct time periods and analyzed crime frequencies across these periods. Initially, we considered breaking the day into broad segments, but after examining the data, we refined these segments as follows:

- **Morning**: 6 AM - 12 PM
- **Noon**: 12 PM - 13 PM
- **Afternoon**: 13 PM - 18 PM
- **Evening**: 18 PM - 22PM
- **Night**: 22 PM - 6 AM



Findings

Contrary to the initial hypothesis, the data shows that crimes are not most frequent during the late night or early morning hours. Instead, crime peaks during:

- **Evening (18 PM - 22 PM)**: The highest number of crimes occurs around 18 PM, which coincides with the evening period.
- **Noon (12 PM)**: There is a significant spike in crime incidents around noon.

The late night hours (22 PM - 6 AM) do show a higher incidence of crimes compared to the early morning hours (6 AM - 12 PM), but they are not the peak periods.

Conclusion of Hypothesis 2

The hypothesis that crimes are more likely to occur during the late night/early morning hours is disproven. Instead, the data reveals that the peak times for crimes are during the evening and around noon. This insight suggests that higher crime rates are associated with periods of high activity rather than low oversight.

Hypothesis 3: Proximity to Crime Category Hotspots Influences Crime Likelihood

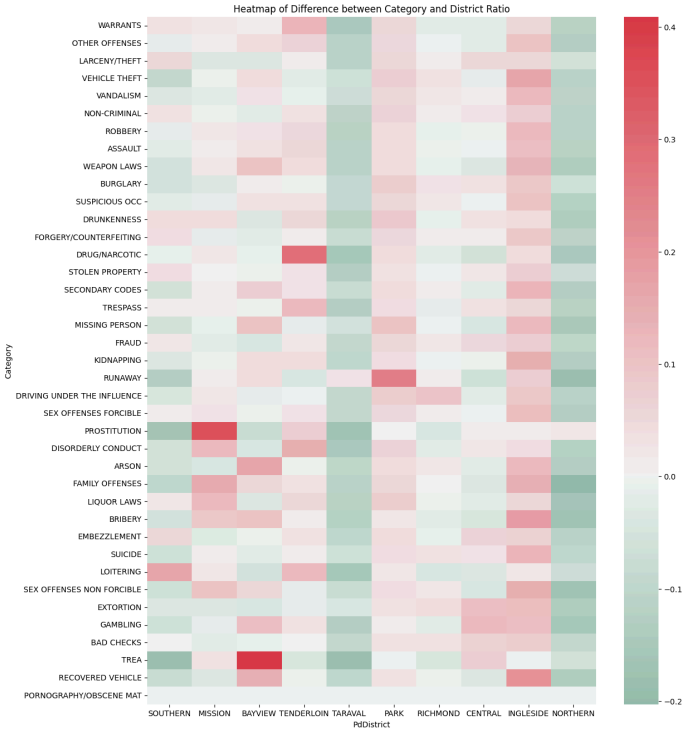
Urban areas often exhibit clustering of similar types of crimes in specific neighborhoods or regions. This hypothesis suggests that the likelihood of a particular crime occurring at a location is influenced by its proximity to crime category hotspots. Hotspots are areas where specific types of crimes are concentrated, and being closer to these areas increases the probability of those crimes happening.

Crime Hotspots and Proximity

The hypothesis is based on the observation that certain areas of a city may be more prone to specific types of crime due to various factors such as socio-economic conditions, the presence of certain types of businesses, or patterns of human activity. If we can identify these hotspots for each crime category and measure the distance from any given location to these hotspots, we can better understand and predict crime occurrences.

Crime Distribution by Police District

To test this hypothesis, we analyzed the distribution of various crime categories across different police districts in San Francisco. We examined whether certain crime types are more prevalent in specific districts compared to others.



Findings

The analysis confirms that the distribution of crimes is not random. Specific types of crimes are significantly more concentrated in certain police districts. For example:

- **Larceny/Theft** is highly concentrated in the Southern district.
- **Drug/Narcotic** incidents are heavily concentrated in the Tenderloin district.
- **Vehicle Theft** is most prevalent in the Ingleside and Bayview districts.

These findings validate the hypothesis that proximity to crime hotspots influences the likelihood of specific types of crimes occurring in those areas.

Conclusion of Hypothesis 3

- Crime distribution in San Francisco shows clear patterns with certain crimes clustering in specific districts.
- Identifying and understanding these hotspots can significantly improve the ability to predict and prevent crimes.
- Police resources and preventive measures can be more effectively allocated by focusing on areas identified as hotspots for specific crime categories.

Dataset Preprocessing Steps

Effective preprocessing is essential for preparing data for machine learning models. We outline the steps taken to preprocess the San Francisco crime dataset, which includes encoding categorical features, handling temporal data, and preparing the data for training and testing.

1. Converting Dates to Timestamps

To handle the temporal aspect of the data, we converted the **'Dates'** column into a numeric format (timestamps), which represents the number of seconds since January 1, 1970.

2. Feature Engineering

We performed several feature engineering steps to prepare the dataset for modeling. These include encoding categorical features and adding new features based on temporal and spatial data.

3. Creating a Preprocessing Pipeline

We utilized ***'ColumnTransformer'*** and ***'Pipeline'*** from scikit-learn to streamline the preprocessing steps. This approach ensures that all necessary transformations are applied consistently to both training and testing datasets.

4. Mapping Encoded Features to Original Categories

For the **'Category'** feature, we need to retain the original category names as we will predict crime categories rather than numerical codes. We create a mapping to revert the encoded values back to their original form.

5. Splitting Data for Training and Testing

We separated the target variable (Category) from the feature set and divided the data into training and testing subsets.

Conclusion

The preprocessing steps outlined above have transformed the raw crime data into a format suitable for machine learning models. Key transformations include encoding categorical variables, handling date-time data, and preparing feature mappings for predictions.

Model Building

The goal of this analysis is to build a model that predicts the category of crime in San Francisco based on various features. We started with a Random Forest classifier and then explored the use of XGBoost to improve accuracy. This report outlines the process and results of building and evaluating these models.

Model Selection

Random Forest Classifier

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. It's known for its robustness and ability to handle large datasets with higher accuracy.

XGBoost (Extreme Gradient Boosting) Classifier

XGBoost is a scalable and efficient implementation of gradient boosting. It provides high performance and accuracy by leveraging gradient boosting frameworks.

Model Performance

- **Random Forest Classifier**: Achieved an accuracy of 26.35% on the test set.
- **XGBoost Classifier**: Improved the accuracy to 27.40%.

While the XGBoost model showed a slight improvement over the Random Forest model, both models struggled to achieve high accuracy due to the complexity and variety of the crime categories.

Conclusion

1. **Temporal and Spatial Patterns**: Crime patterns in San Francisco are influenced by both temporal and spatial factors. Weekdays and specific districts have higher crime rates, and certain crime types cluster in particular areas.
2. **Model Performance**: While the Random Forest and XGBoost models provided a starting point, their performance suggests room for improvement. Further feature engineering, such as incorporating resolution outcomes or more granular spatial data, could enhance prediction accuracy.
3. **Feature Engineering**: The creation of time-based features and encoding of categorical variables are critical steps that help models leverage temporal and spatial dimensions of the data.
4. **Predictive Modeling**: Advanced techniques, including hyperparameter tuning and addressing class imbalance, should be explored to improve model performance. Incorporating additional data sources, like socio-economic factors, could also provide deeper insights.

Future Work

To enhance the predictive capabilities and understanding of crime dynamics in San Francisco, future work could focus on:

Enhanced Feature Engineering

- **Incorporate Resolution Outcomes**: Including data on whether crimes were solved or not could provide additional context for predictive modeling.
- **Temporal Aggregation**: Creating aggregated features such as rolling averages or crime counts over different time windows (e.g., last week, last month) could capture trends and periodicity in crime incidents.

Attend useful hypothesis into training model

Hypothesis 3 offers useful information which can understand crime hotspots and proximity is not random. We may add it as a new feature into our model in order to give the prediction better accuracy.

The analysis of San Francisco crime data provided valuable insights into the temporal and spatial patterns of crime. While the initial models demonstrated the complexity of predicting crime categories, they also highlighted areas for potential improvement through advanced feature engineering and attending useful hypotheses into training models.

Reference

We reference source code of **feature engineering** and **how to encoding data** from Kaggle.

https://ieeexplore.ieee.org/document/8768367

https://www.kaggle.com/competitions/sf-crime

Contribution

黃鈺凱, 109370210 50%
陳國誌, 109370211 50%

All resources are in our Github.

https://github.com/edwall201/SF-CriminalRate-Prediction