San Francisco Crime Classification

Jait Purohit, Saumya Shah

Computer Science Department, California State University, Sacramento
Sacramento
jaitpurohit@csus.edu
saumyashah@csus.edu

Abstract— **Public safety and protection from crime are important for better development of the city. Every city is trying to mitigate these types of crimes and to help the civilians. This document examines, evaluates and predicts patterns of crime existing in San Francisco and its suburbs according to the dataset provided by Kaggle data-challenge. As a comprehensive process, we have divided the project into exploratory data analysis phase, feature-engineering, pre-processing of data, model training and prediction, and result analysis. For better improvement in results, we are trying different combinations of data, classifying it in more meaningful manner, so as to improve the accuracy of the evaluation models. Our solution is more focused on generalised area rather than specific locations, which helped us to get better results. Finally, we predict and compare the results from different classifiers and dwell into improvements for future work.**

Keywords— **SF crime data, Kaggle, bay-area crime, multi-label classification, deep learning, machine learning**

I. INTRODUCTION

Crimes these days are getting miserable in big cities, which are affecting civilians. San Francisco is ranked in amongst top cities in United States affected by various types of crimes. The solution to this problem statement is important because every citizen has the right to stay in a safe-environment and live a fruitful life. San Francisco Police Department has released crime-related data from 2003-2018 available to general public. Taking the data into account and understanding patterns in the crime dataset has helped in the field of Machine Learning and Artificial Intelligence. These algorithms are fast and efficient, providing close to accurate results, by doing proper data-research.

In our project, we are using San Francisco crime dataset uploaded on Kaggle data-challenge to predict which category of crime is most likely to take place, at a given location(place), and time. The feature inputs of our data are date – (broken into day, hour, minutes, month and year), location – (based on address of the areas in city). The output feature is the category of crime that is supposed to occur. We have done feature engineering on date, place related columns in the dataset and ranked according to its importance. We have used encoding schemes likes label-encoding, one-hot encoding techniques in our data-processing.

- Exploratory data analysis of training data
- One-hot encoding of categorical input features
- Dividing 'address' column into streets, blocks and intersection, to have more meaningful location information
- Label encoding of categorical crimes
- Feature important analysis
- Implemented logistic regression, SVM, Nearest neighbour, random forest, deep-learning neural networks models like artificial neural network and convolutional neural networks.
- New classifiers like xgboost and adaboost

Further in this paper, we are briefly discussing the problem statement, algorithm design and its structure. Moreover, we are evaluating experimental results. Current research and related work have also been added. The conclusion of our research paper has been given in a nutshell. Finally, work distribution and learning experience have been briefly discussed below.

II. PROBLEM FORMULATION

The problem has been formulated in most precise manner and can be explained in distinct parts.

1. Exploratory data analysis of crime:
    - We have plotted different graphs about crimes in different location of the city
    - We have plotted graphs about crime categories to analyse top five crimes
    - We have done analysis on crime and its time of occurrence.

2. Feature engineering and Data Pre-processing
    - Encoding schemes like one-hot encoding and label encoding for converting categorical data into numerical data
    - Checking of null values
    - Removing un-necessary rows

3. Our inputs features are location related data, time related data, police district data. Our output is category of crime type. There are 9 fields in original training data.

    Dates - timestamp of the crime taken place
    Category - category of the crime type
    Address - the street address of the crime location
    X - Longitude coordinates
    Y – Latitude coordinates
    Descript - description of the crime incident
    DayOfWeek - day of the week
    PdDistrict - Police Department District
    Resolution - crime incident result

4. Building prediction models
    - Classification models like logistic regression, nearest neighbour, svm, gaussian naïve bayes, random forest
    - Deep learning neural networks models like artificial neural network and cnn are used her
    - New models like xgboost and adaboost are also used to predict type of crime

We are using multi-label classification problem to predict crime types using above mentioned models. Different algorithms mentioned above gives different accuracy. Finally, we are using our best model to predict on test data.

III. System/Algorithm Design

This section entails our multi-label classification algorithms defined in the field of machine learning and artificial intelligence, which helps to predict crime type based on input feature set.

A. System Architecture

Based on exploratory data analysis and doing proper feature engineering, we are deriving lot of important features based on the given raw training data. For example, address field is divided into street1 and street2 types, which defines whether the address is an intersection, is a block or not. We are also doing one-hot encoding of these input categorical features to convert into numerical column values. Furthermore, date field is divided into hours, minutes, days, and years. These are helpful in defining the seasons based on month column, defining hour sections based on hour column and classifying whether it is a weekday or weekend based on 'day of week' column. They are also one-hot encoded. These important fields become our input feature set and help to predict our output feature, i.e. category (crime) column, which is label encoded as it is categorical data.
Feature important analysis is done here using input perturbation technique, which lists the rank of input features.
Finally, we are using using classification algorithms to predict multi-label classification of crime types.

B. Modules

3.1) Basic Classification algorithms

1) Logistic Regression: Logistic regression is classification algorithm used for predicting class membership based on probabilities. It accounts for probabilities beyond 1.

$$\text{logit}(\pi(x))=\log(\pi(x)1-\pi(x))=\beta 0+\beta x$$

The logistic regression model uses logistic cumulative distribution function(cdf). For example, whether the tumor is malignant or not can be predicted by logistic regression.

2) Nearest Neighbour: K-nearest neighbour is used for easy of interpretation and low calculation time.
Pseudo code for knn can be given as:
a) load the data and initialise the value of k
b) for getting predicted class, evaluate the distance between training data and test data, and sort the values based on its distance.
c) get top k rows and get the most frequent class based on these rows
d) return the predicted class

Nearest neighbour algorithm can be used in political science to classifying a voter, whether he will "vote" or "will not vote", or to "vote democrat" or "vote republican". Here, knn algorithm is based on feature similarity.

3) Support Vector Machines: SVM is a supervised algorithm. Here, we are finding the linear hyperplane that differentiates between classes of data points. Hyper-planes are selected based on the optimality of margin value. Our aim will be to maximise the margin, and thus that will be the best hyper-plane possible.

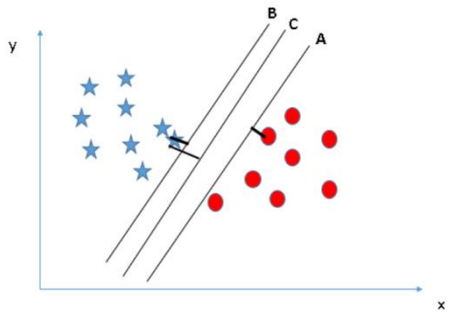

Fig. 1 support vector machine algorithm

4) Random Forest: Random Forest works on the concept of decision trees, which are the basic building blocks of this algorithm. It is a supervised learning algorithm, which is used to create a forest. The more the number of trees, the more the accurate our result for the model. Random forest solves the problem of overfitting and handle missing values and can be used to model categorical data.
In our project, we are getting best result in terms of accuracy for random forest amongst our basic models.

3.2) Deep Learning Neural Networks algorithms

1) Artificial neural networks: This algorithm is a part of neural network which consists of input and output layers. It has many fully-connected hidden layers of neurons, which learn the weight based on input features. They have an important concept of

back-propagation which helps in adjusting the weights to produce better outcome. Activation functions like tanh, relu and sigmoid are used here to adjust the weights. Optimizers like rmsprop, adam, sgd also used for hypertuning of parameters.

2)

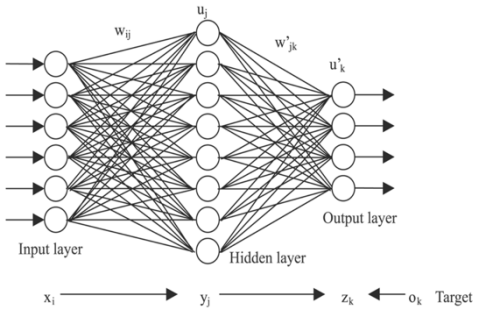

Fig. 2 Artificial Neural Network Architecture

This neural network can be used for making cars drive autonomously on roads and reading our minds.

In our project, we are adjusting weights using different activation functions like tanh, relu, sigmoid while trying different optimizers like adam, rmprop, sgd and different neuron counts to predict crime type category based on input feature set.

3) Convolutional Neural networks: CNN or Convnets is one of the main categories to do image recognition, image classification. They are also used in object detections and recognition of faces. CNN takes our input features as image of array of pixels and predict output. It has three layers defined:

a) Convolutional stage
b) Non-linearity stage
c) Pooling stage

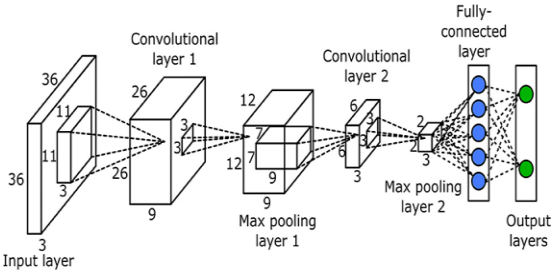

Fig. 3 Convolutional Neural Network

3.3) New Classifiers

1) Adaboost Classifier: Adaboost is similar to random forest classifier. They retrains the algorithm in iterations based on accuracy of previous training data. They give more correct and precise results since it depends on weak classifier for its decision. One such application for Adaboost classifiers is face recognition system. Moreover, each instance of training data is given a weight. Predictions are made using the weighted average of weak classifiers.

2) XGBoost Classifier: XGBoost means eXtreme Gradient Boosting. This algorithm is used for better execution speed and model performance. They grow their decision-trees in level-wise manner.

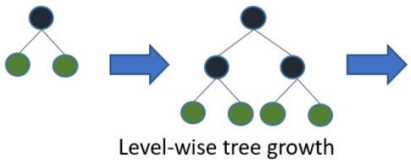

Fig. 4 Level-wise tree growth in XGBoost

IV. EXPERIMENTAL EVALUATION

1) Methodology:

- San Francisco Crime Classification from Kaggle data-challenge dataset was used in our project. We split the data into 80% training data and tested on rest 20%.
- Experimental setting done was on the data, to remove the redundant and less-important data. For example, we have gotten rid of latitude and longitude columns as we have address, which is more relevant. Date column was split into year, month, day, hour, minutes, to have more meaningful classification of the crime related data. Few outliers were also handled. Encoding schemes were also used for conversion purposes. Feature important analysis was done using input perturbation approach.
- Classification evaluation measures like accuracy, confident matrix, precision-recall were used between model comparison.
- Methods like logistic regression, nearest neighbour, support vector machines, gaussian naïve bayes, random forest, artificial neural networks, convolutional neural network, adaboost classifiers and xgboost classifiers were implemented and compared.

2) Results:

Baseline Algorithms	**Metrics**			
	Accuracy	**Precision**	**Recall**	**F1 Score**
Logistic Regression	45.13%	.4233	.4513	.4122
KNN	37.62%	.3569	.3762	.3585
SVM	43.35%	.4031	.4335	.3940
Gaussian NB	41.76%	.3920	.4176	.3661
Random Forest	44.99%	.4353	.4496	.4375

Deep Learning Models	Metrics			
	Accuracy	**Precision**	**Recall**	**F1 Score**
Artificial Neural Network	46.93%	.4453	.4693	.4334
Convolutional Neural Network	45.59%	.4248	.4559	.4177

New Models	Metrics			
	Accuracy	**Precision**	**Recall**	**F1 Score**
AdaBoost Classifier	43.58%	.4053	.4358	.3975
XGBoost Classifier	43.58%	.4053	.4358	.3975

Our best model achieved was artificial neural network. Even though, logistic regression and random forest classifier were close enough to our best model achieved, artificial neural network classifier is the optimal, theoretically and practically. The ability to adjust weights using back-propogation in artificial neural network has given this best model, which is not the case in logistic regression. Training support vector machines and convolutional neural network are considerably giving less accurate results and take longer time to execute. Comparison with new models implemented, only suggested that adaboost classifier was closer to artificial neural network, but artificial neural network can substantially solve the problem of overfitting here. So, with an accuracy score of 46.93% outperforms other classifiers for crime classification.

V. Related Work

Crime analysis and its prediction revolves around the analysis of spatial and temporal data. Comparison of crimes in different seasons has been depicted by the authors of [1]. They were able to prove their hypothesis using Negative Binomial Regression and Ordinary Least Squares (OLS).

Comparison analysis based on day of the week was carried by authors of [2] between different cities. As an added feature, the safest and notorious streets were also analyzed in this research paper.

Venturini *et al.* [3], in their research paper have observed different seasonal patterns in crime and has analysed the type of crime changes with the month. They have used Lom-Scargle periodgram [4] to achieve this. The AstroML Python package was also used here.

Cohn, Ellen G. [5] in their research work "Weather and crime" has pointed out strong coherence between number of days with minimum or maximum temperature and the robbery rates in the city.

Ensemble Learning can improve accuracy of various ML tasks, as it combines many models using voting. This was presented in our of the research articles from Kaggle ensembling guide [6].

VI. Conclusion

To conclude in a nutshell, this project has been a challenge and learning curve, right from exploratory data analysis phase, to data-preprocessing based on feature engineering and applying the training data to train different classifiers.

Exploratory data analysis and feature extraction and its engineering were mainly done in parallel. As time related information was easy to extract as features, more analysis on location-based features was taken into account. Location information from address was split into multiple columns to have more meaningful extraction of feature-set, which in turn will help in improving the accuracy of models. For example, address location of the crime was split based on intersection of two streets, blocks.

Based on our training data, we have trained some basic models like logistic regression, nearest neighbor, support vector machines, gaussian naïve bayes and random forest. We have calculated different evaluation measures like accuracy, precision, recall and F1-score. Deep learning neural network models like artificial neural network and convolutional neural network were also trained. We have implemented new models in this research paper like adaboost and xgboost classifier. Artificial neural network outperforms other classifiers in terms of accuracy, as this was our best model for multi-class classification scenario.

As a future work on this topic, principle component analysis (PCA) for dimensionality reduction may improve performance. Moreover, we can use twitter feeds of a particular area, to make use of sentimental analysis to observe relations of sentiment of the civilians and crime rate.

Additionally, this same approach was done on Chicago crime data-set, to compare crimes in San Francisco and Chicago.

VII. Work Division

- Phase 1: Exploratory data-analysis includes plotting of graphs and understanding various features and having co-relation between them was done as a team.
- Phase 2: Feature Engineering, label-encoding of category column, one-hot encoding of categorical input features was done as a team.
- Phase 3: Feature importance analysis using input perturbation technique was done a team.
- Phase 4: Model training and evaluation was done as a team.

VIII. Learning Experience

Exploratory data analysis has given us insights many data-visualization methods and approaches to think in the direction of feature-engineering phase. Feature importance analysis phase taught us about input perturbation technique. Additionally, we learnt model training, evaluation strategies and its comparison. New models like random forest, adaboost classifier and xgboost classifiers were also implemented. Overall, it was a great learning curve, overcoming the challenges as a team.

IX. REFERENCES

[1] S. J. Linning, M. A. Andresen, and P. J. Brantingham, “Crime seasonality: Examining the temporal fluctuations of property crime in cities with varying climates,” International journal of offender therapy and comparative criminology, vol. 61, no. 16, pp. 1866–1891, 2017.

[2] T. Almanie, R. Mirza, and E. Lor, “Crime prediction based on crime types and using spatial and temporal criminal hotspots,” arXiv preprint arXiv:1508.02050, 2015.

[3] L. Venturini and E. Baralis, “A spectral analysis of crimes in san francisco,” in Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. ACM, 2016, p. 4.

[4] N. R. Lomb, “Least-squares frequency analysis of unequally spaced data,” Astrophysics and space science, vol. 39, no. 2, pp. 447–462, 1976.

[5] Cohn, Ellen G. "Weather and crime." British journal of criminology 30.1 (1990): 51-64.

[6] https://mlwave.com/kaggle-ensembling-guide/

[7] “Crime Prediction and Classification in San Francisco City” by Addarsh Chandrasekar, Abhilash Sunder Raj and Poorna Kumar, Stanford.

[8] Exploratory Data Analysis And Crime Prediction In San Francisco by Isha Pradhan, San Jose.