Predicting neighborhoods

with exclusively high risk of credit card fraud

Wen-Chieh Sung

1. Introduction:

1.1. Background:

As economy grow and online payment become prevailing, credit card fraud also finds its way to expand. Theft by use of or counterfeit of credit card, for example, are very common forms of credit card fraud. However, instead of a one-sided game, markets around the world also find their way to fight back. Unfortunately, the winning in some countries leads to the misfortune of their surrounding countries as the case in Europe. According to the report of FICO (a company providing card alert service) in 2017, drop in the successful fraud rate in UK and France has possibly led to the rise of success fraud rates in Austria, Denmark, Norway, Sweden, Poland and Russia. Therefore, capability to predict the neighborhoods with high risk of credit card fraud should be very helpful for financial industries and government of these vulnerable countries to fight against the upcoming fraud.

1.2. Problem

Primary venues of a neighborhood might have great influence on the risk of credit card fraud. For example, neighborhoods with many financial facilities or shopping places could be fraudsters' favorite spots. Therefore, here we aim to build a model to predict the risk of credit card fraud of neighborhoods based on data of venues scraped from the Foursquare API. In this project, we will train the model with San Francisco data, and evaluate the model performance with Chicago data.

1.3. Interest

Successfully predicting the neighborhoods with high risk of credit card fraud could help local government get prepared in advance to protect citizen from the risk. For financial institutions, they could invest more effort in monitoring transactions within the neighborhoods of exclusively high risk to enhance their efficiency of detecting credit card fraud.

2. Data acquisition and cleaning

2.1. Data sources

Crime data:

We downloaded the crime dataset of most recent year of San Francisco and Chicago from DataSF (here) and DATA.GOV (here) respectively. The two websites provide constantly updated crime data recorded by local police departments. These data include the coordinates of police station, date of incident, crime description, etc.

Neighborhoods json file:

For San Francisco data, we used the Analysis Neighborhood defined for the purpose of providing consistency in the analysis and reporting of socio-economic, demographic, and environmental data, and data on City-funded programs and services, whose json file can be downloaded from here.

For Chicago data, we downloaded the json file from here, which is a repository on Github that records the method to transform coordinates into according neighborhood name in Chicago.

Venues data:

The venues data were generated from the Foursquare API. The data include the popular spots and their categories in each neighborhood.

2.2. Data wrangling

According to our goal, we only need the categories of crime, description of crime and coordinates of the police station where the incident was reported to. These data correspond to the column 'Category', 'Descript', 'Y' and 'X' in San Francisco's crime data, and 'PRIMARY DESCRIPTION', 'SECONDARY DESCRIPTION', 'LATITUDE', and 'LONGITUDE' in Chicago crime data. Therefore, we extract these columns and drop the rest.

The only clues to check if a crime belongs to credit card fraud are its category and description. We found that credit card fraud should belong to 'FRAUD' and 'DECEPTIVE PRACTICE' in San Francisco and Chicago crime data respectively. Afterwards, we look further into their criminal descriptions and found the most related descriptions are 'CREDIT CARD, THEFT BY USE OF' and 'CREDIT CARD, THEFT OF' for San Francisco data and 'CREDIT CARD FRAUD' for Chicago data. After dropping the data with invalid coordinates (i.e., incomplete or not located in San Francisco or Chicago), there are 23068 and 4537 credit card fraud crimes in San Francisco and Chicago respectively.

To find out the neighborhoods with exclusively high risk of credit card fraud, we calculated the credit card fraud rates of each neighborhood as number of credit card fraud of a neighborhood divided by the total number of credit card fraud in San Francisco or Chicago. We then drew a choropleth of San Francisco data and found that some neighborhoods did have exclusively higher credit card fraud rates than others (Fig. 1). We labeled the outliers of credit card fraud rates of neighborhoods as ‘outliers’ and the other as ‘average’. There are 4 (out of 41) and 6 (out of 78) neighborhoods labeled as outliers for San Francisco and Chicago respectively (Fig. 2). The descriptive statistics of credit card fraud rates of neighborhoods are shown in Table 1.

We used the explore function of Foursquare API to get the top 100 popular venues of each neighborhood within the radius of 500 m from the centroid of the neighborhood. Because the categories returned by the explore function could be too specific to serve as the feature of a model of broad use. Therefore, we looked up the primary category of each category, and replace the original categories with their primary categories. According to the Foursquare website, there are 10 primary categories in total, including ‘Arts & Entertainment’, ‘College & University’, ‘Event’, ‘Food’, ‘Nightlife Spot’, ‘Outdoors & Recreation’, ‘Professional & Other Places’, ‘Residence’, ‘Shop & Service’, ‘Travel & Transport’. Lastly, we used the venues data of neighborhoods to calculate the percentage of each primary category of each neighborhood. The metric of San Francisco is then used for training model (Fig. 3), and the metric of Chicago is used for evaluating the performance of the model.

3. Building classification model and model evaluation

3.1. Classifier and evaluation criteria

In this project, we aim to predict the outliers of credit card fraud rates of neighborhoods. We’ll use the classifier we’ve learned so far to do the prediction, that is, K nearest neighborhoods (KNN), decision tree, support vector machine (SVM) and logistic regression. These algorithms are done by functions from scikit-learn library.

Underestimating the risk of credit card fraud could be devastating to financial institution and public wealth, therefore, high value of recall of a model is very desirable. Thus, we’ll evaluate the performance of a model first by its recall of outliers, secondly precision of outliers. The adjustment of model parameters like ‘n_neighbors’ in KNN model and ‘max_depth’ in decision tree model is also based on the criteria.

3.2. Resampling of San Francisco data (training set)

The training data (San Francisco) are very imbalanced as the neighborhoods labeled as outliers are far less than the neighborhoods labeled as average. The classifier we’re going to use have a bias towards classes which have major number of instances. They tend to only predict the major class data successfully. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class, which is not desirable in our case.

Resampling techniques are widely applied to balance imbalanced class. The main objective of balancing classes is to either increasing the frequency of the minority class (i.e., over-sampling) or decreasing the frequency of the majority class (i.e., under-sampling). This is done in order to obtain approximately the same number of instances for both the classes. In our case, we don't consider to under-sample the San Francisco data because it can discard potentially useful information which could be important for building rule classifiers especially with this low number of sample (41 in total). Therefore, we'll apply over-sampling to balance the class of San Francisco data.

In our project we used two types of over-sampling techniques. The first is naive over-sampling, which increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample. However, it could increase the likelihood of overfitting since it replicates the minority class instances. The second is synthetic minority over-sampling technique, or SMOTE, which takes a subset of data from the minority class as an example and creates new synthetic similar instances. This technique mitigates the problem of overfitting caused by naive oversampling as synthetic instances are generated rather than replication of instances.

4. Results and discussion

Overall, our models do not perform well. If we accept a little loss in recall, let's say recall = 0.83 at minimum (i.e., fail to predict one outliers of Chicago data). Then the qualified models will be SVM model and logistic regression model trained by naive oversampled data, and decision tree model, SVM model and logistic regression model trained by SMOTE oversampled data (Table 2). However, none of these models had satisfactory precision of predicting outliers (Table 2). High recall of these models came at the cost of precision, which led to many false positive in prediction. Among the qualified models, the highest precision is 0.10 from the decision tree model trained by SMOTE oversampled data, meaning that there's only one neighborhood actually at high risk of credit card fraud out of ten neighborhoods predicted to be at high risk (Table 2; Fig. 5b). Although there's no relevant model for us to compare the precision of prediction, we find a report stated that only 1 in 5 transactions declared as fraud by fraud detection model was truly fraud. Therefore, our models are not even close to this level of precision.

We go through the data we used and found out the following possible reasons responsible for the low precision of the model:

1. **Indistinct definition of credit card fraud:** As I mentioned in the data section, the only clue we can define a crime as credit card fraud is through the category and description of each crime. However, compared to San Francisco data, the criminal description of Chicago data is not very informative for credit card fraud, as the description was just 'credit card fraud'. The term credit card fraud could stand for various form, which might not be equivalent to those of San Francisco data, contributing to the bad prediction of our models.
2. **Small number of training data:** Small number of training data enhance the dilemma of overfitting and underfitting. Since the outliers of San Francisco data are few, split the data into training and test set could easily result in underfitting of the model. Therefore, we use the whole San Francisco data as training set. However, as we have seen, this might result in overfitting of the model. We can infer the overfitting from the overall bad recall of the KNN models (0.50 from naive oversampling and 0.33 from SMOTE oversampling). The credit card fraud rates of outlier meighborhoods in San Francisco have narrower range and higher value than those in Chicago (Fig. 1), therefore the outlier neighborhoods in Chicago with lower rates are not 'near' neighbors to the outlier neighborhoods in San Francisco.

5. Future work

Honestly, we have little to adjust for the problem of indistinct definition of credit card fraud. At best, we can contact the data owner to see if we can get more detailed description about the credit card fraud of Chicago data. For the problem of small number of training data, we can try to collect the whole credit card fraud data of USA to train the model, and use it to predict the risk of credit card fraud of neighborhoods in other country (e.g., Canada) to see if the larger amount of raw data solves the dilemma of overfitting and underfitting.

Tables:

Table 1. Descriptive statistics of credit card fraud rates of neighborhoods in San Francisco and Chicago (unit: %).

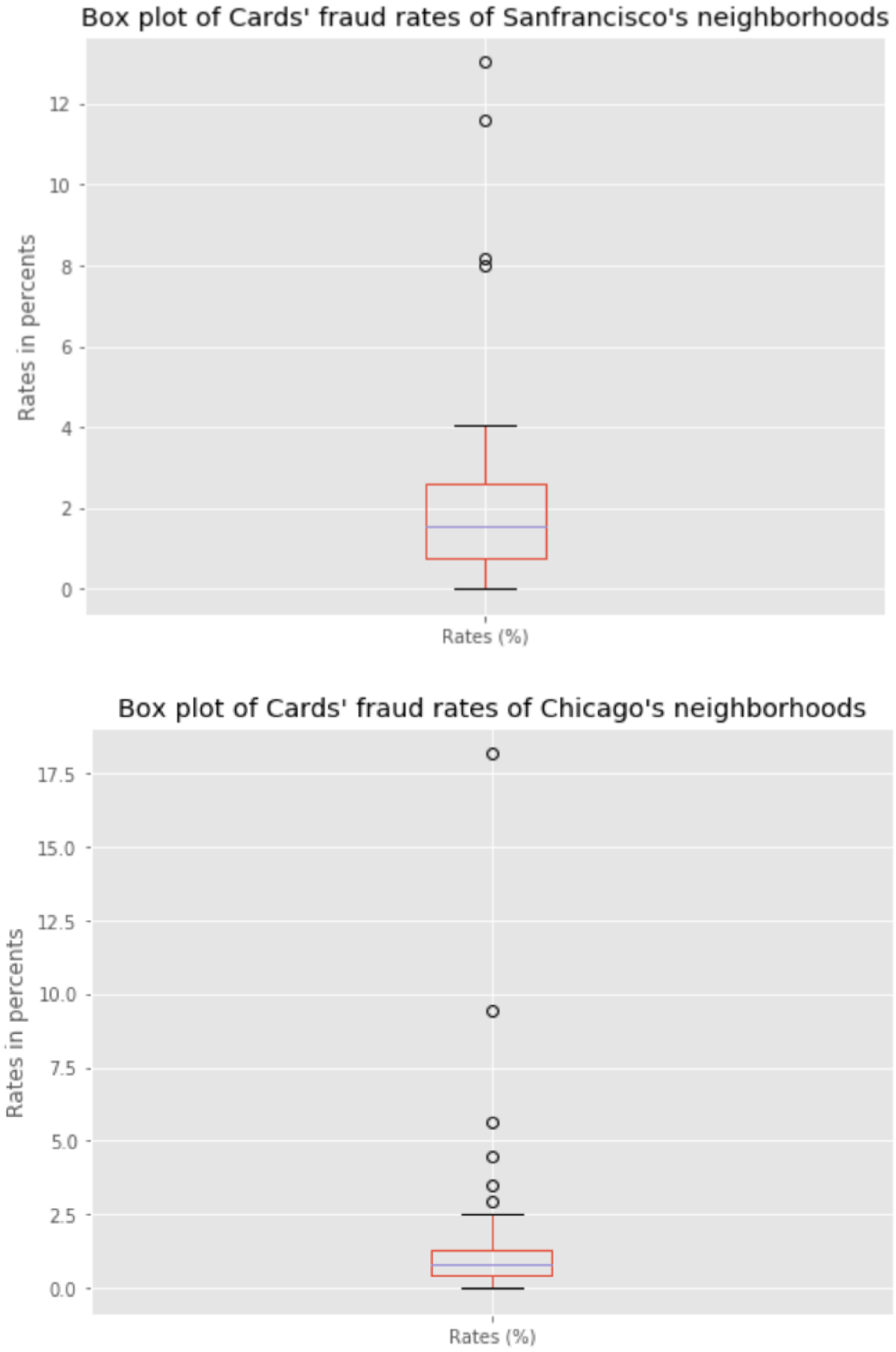
	San Francisco	Chicago
Mean	2.44	1.28
Std	2.88	2,36
Min	0.01	0.00
25% quantile	0.76	0.40
50% quantile	1.54	0.77
75% quantile	2.62	1.27
Max	13.03	18.17

Table 2. Recall and precision of outliers and weighted average fi-score of each model.

Naive	**KNN**	**Decision Tree**	**SVM**	**Logistic Regression**
Recall	0.5	0.67	1.00	0.83
Precision	0.08	0.12	0.09	0.08
f1-score	0.63	0.69	0.26	0.36
SMOTE	**KNN**	**Decision Tree**	**SVM**	**Logistic Regression**
Recall	0.33	0.83	1.00	1.00
Precision	0.07	0.1	0.09	0.09
f1-score	0.70	0.52	0.18	0.26

Figures:

Fig 1. Boxplot of credit card fraud rates of neighborhoods in (a.) San Francisco and (b.) Chicago.



Box plot of Cards' fraud rates of Sanfrancisco's neighborhoods

Box plot of Cards' fraud rates of Chicago's neighborhoods

Fig 2. Choropleth of credit card fraud rates of neighborhoods in San Francisco.

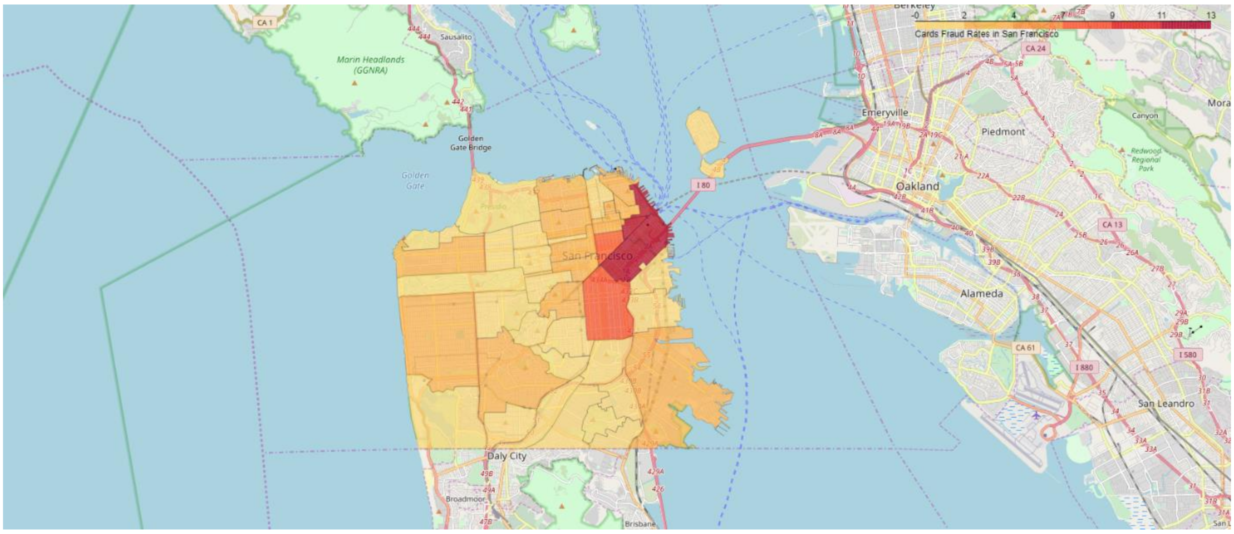
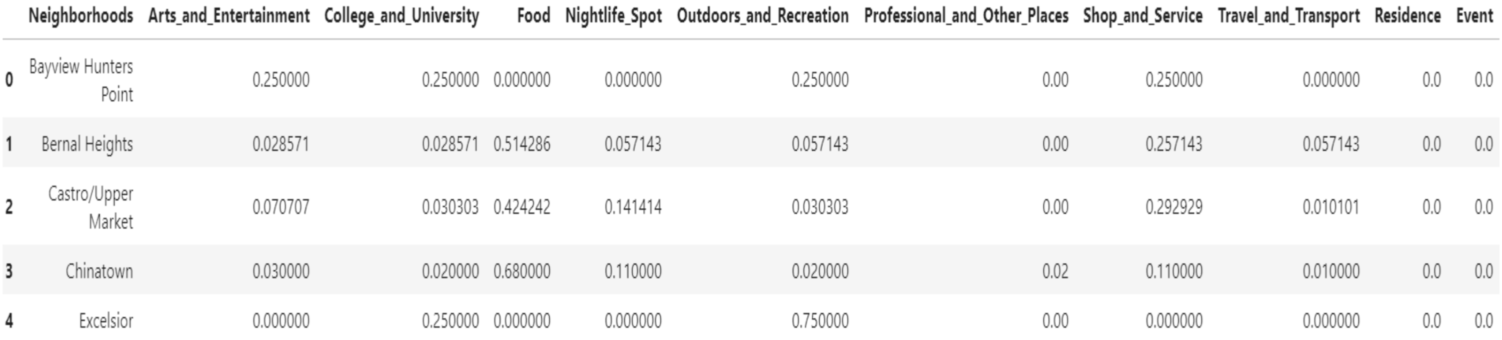


Fig. 3 A demonstration of the metric of training data.



	Neighborhoods	Arts_and_Entertainment	College_and_University	Food	Nightlife_Spot	Outdoors_and_Recreation	Professional_and_Other_Places	Shop_and_Service	Travel_and_Transport	Residence	Event
0	Bayview Hunters Point	0.250000	0.250000	0.000000	0.000000	0.250000	0.00	0.250000	0.000000	0.0	0.0
1	Bernal Heights	0.028571	0.028571	0.514286	0.057143	0.057143	0.00	0.257143	0.057143	0.0	0.0
2	Castro/Upper Market	0.070707	0.030303	0.424242	0.141414	0.030303	0.00	0.292929	0.010101	0.0	0.0
3	Chinatown	0.030000	0.020000	0.680000	0.110000	0.020000	0.02	0.110000	0.010000	0.0	0.0
4	Excelsior	0.000000	0.250000	0.000000	0.000000	0.750000	0.00	0.000000	0.000000	0.0	0.0

Fig. 4 Heat-map of confusion matrix of models trained by naive over-sampled San Francisco data. (a) KNN (b) Decision tree (c) SVM (d) Logistic regression

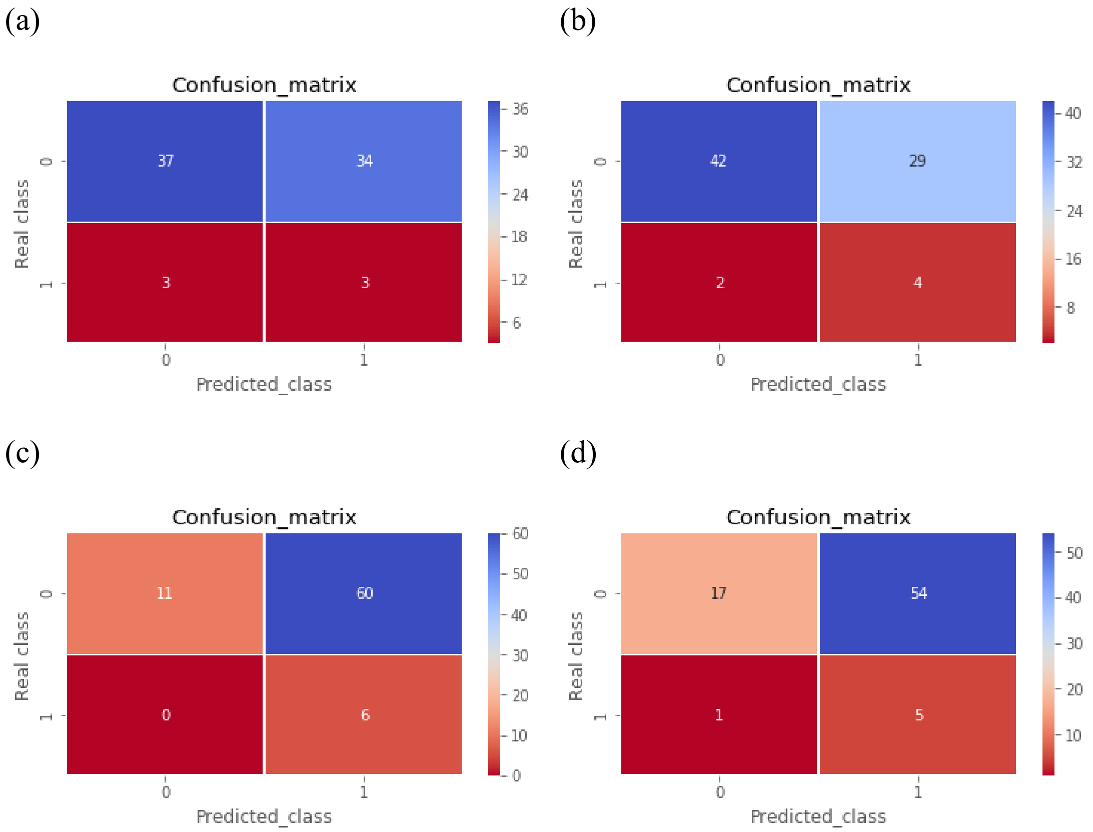
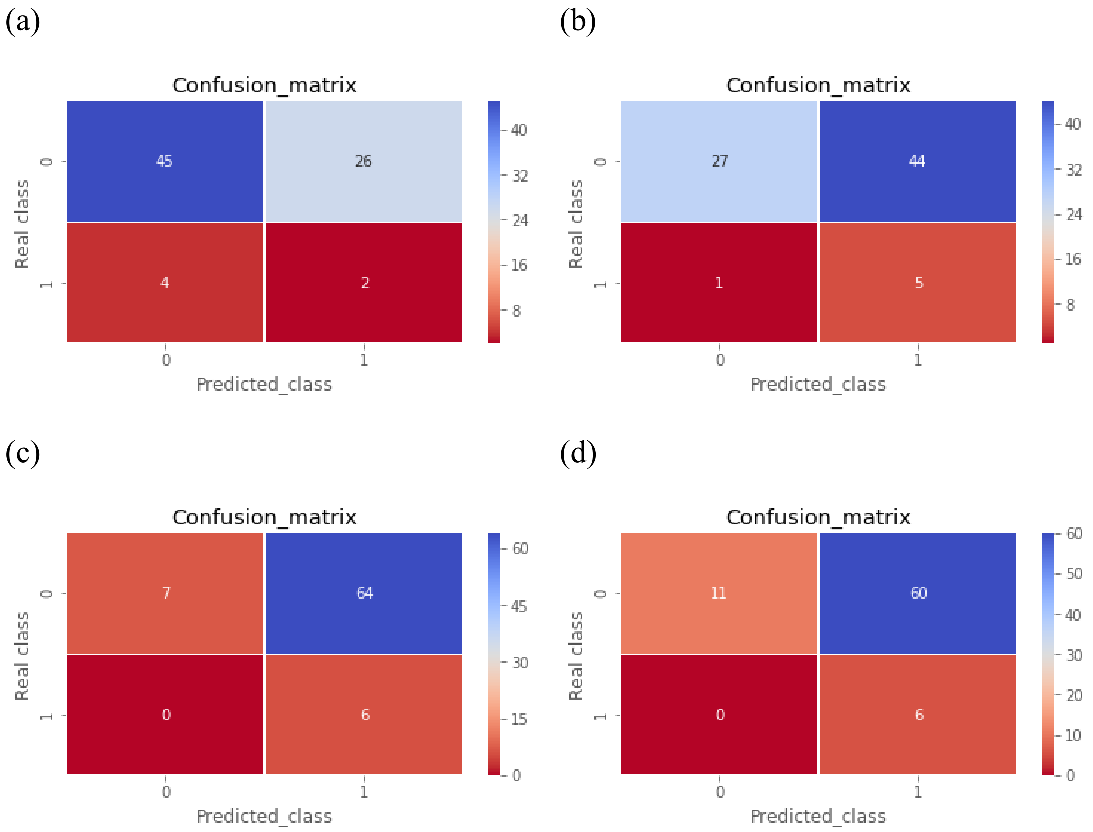


Fig. 5 Heat-map of confusion matrix of models trained by SMOTE over-sampled San Francisco data. (a) KNN (b) Decision tree (c) SVM (d) Logistic regression


(a)
Confusion_matrix
Real class
Predicted_class
(b)
Confusion_matrix
Real class
Predicted_class
(c)
Confusion_matrix
Real class
Predicted_class
(d)
Confusion_matrix
Real class
Predicted_class