Predicting neighborhoods with exclusively high risk of credit card fraud

Coursera capstone project

Wen-Chieh Sung

Created by TNS
from Noun Project

Created by john melven
from Noun Project

Predicting neighborhoods with high risk of credit card fraud is valuable

- **Background**: credit card fraud expand geographically as growth of economy.

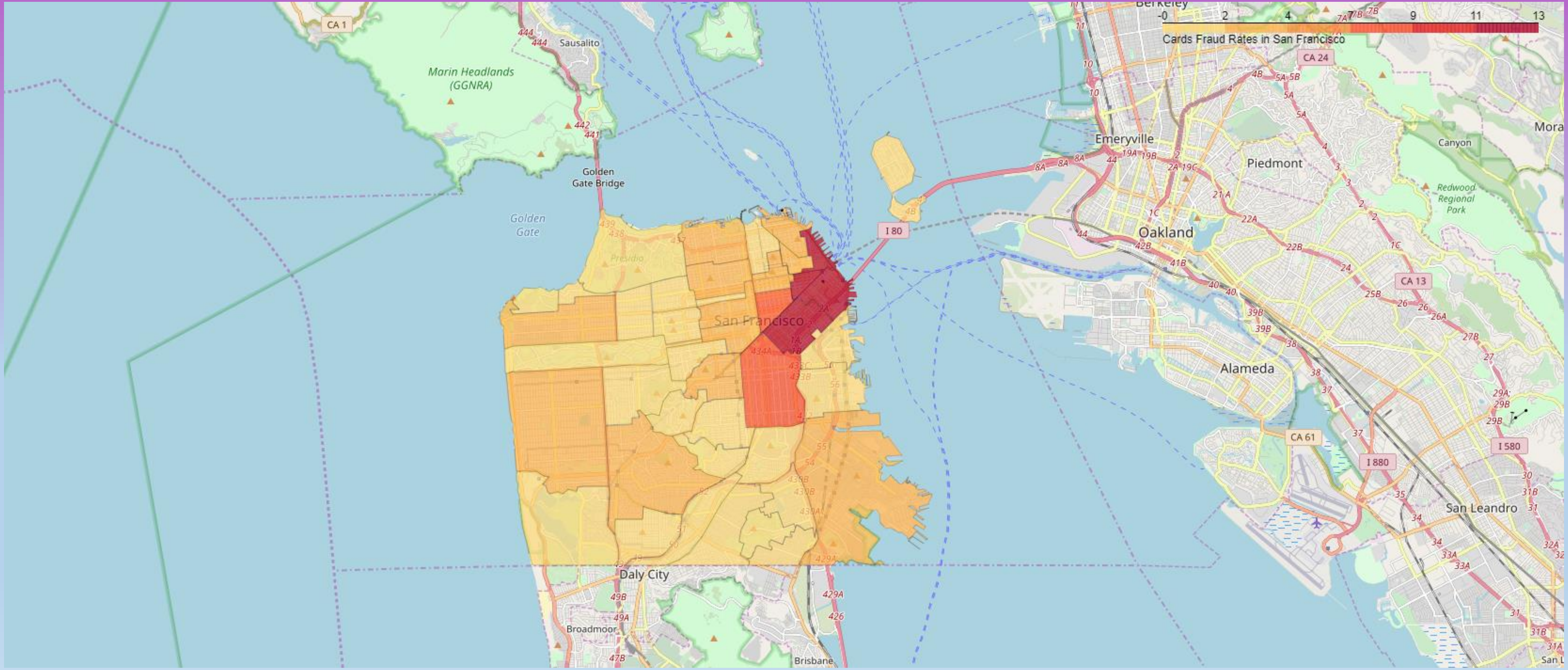
- **Problems**: venues data of neighborhoods may help in predicting neighborhoods with high risk of credit card fraud.

- **Interest**:
  - ✓Government: get prepared in advance to protect citizen from the risk
  - ✓Financial institution: enhance their efficiency of detecting credit card fraud

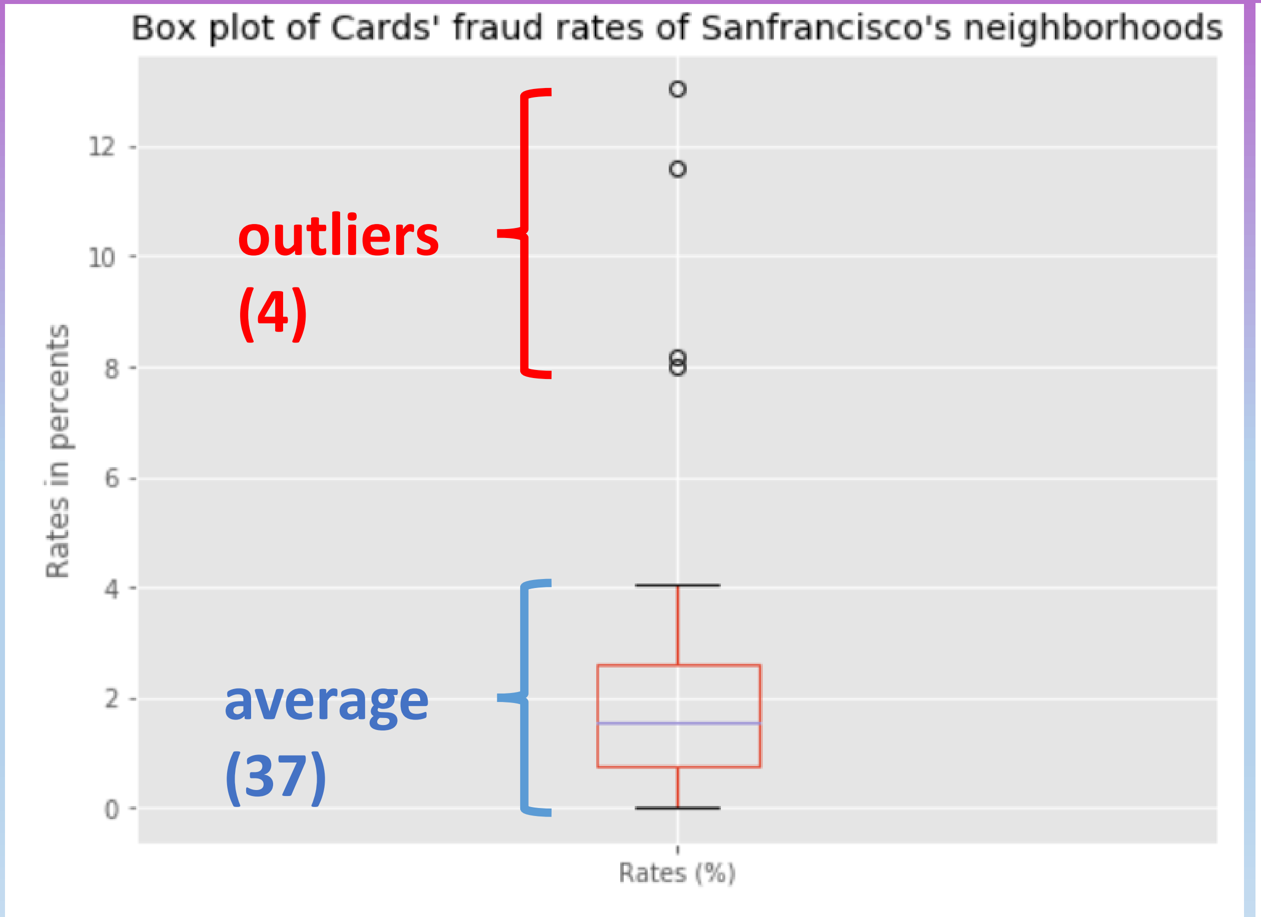
Data acquisition and cleaning

- San Francisco and Chicago crime data of the most recent year are downloaded from DataSF and DATA.GOV.
- We extract criminal incidents with description of ‘CREDIT CARD, THEFT BY USE OF’ and ‘CREDIT CARD, THEFT OF’ of San Francisco data, and ‘CREDIT CARD FRAUD’ of Chicago data as credit card fraud. There are 23068 and 4537 incidents in San Francisco and Chicago respectively.
- San Francisco data are used as training set, while Chicago data are used as test set

Some neighborhoods are in high risk (San Francisco)

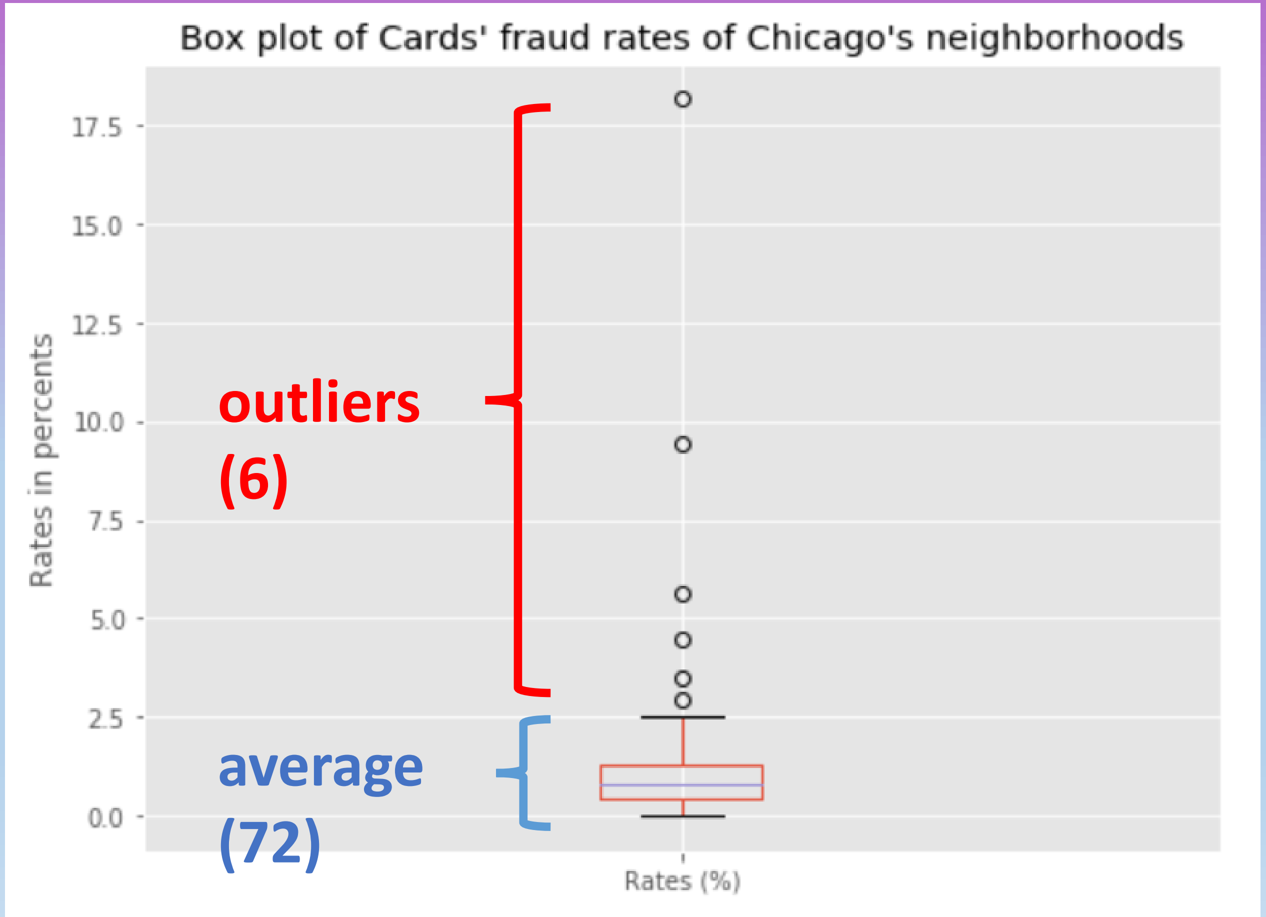


Categorizing data into ‘outliers’ & ‘average’



outliers (4)

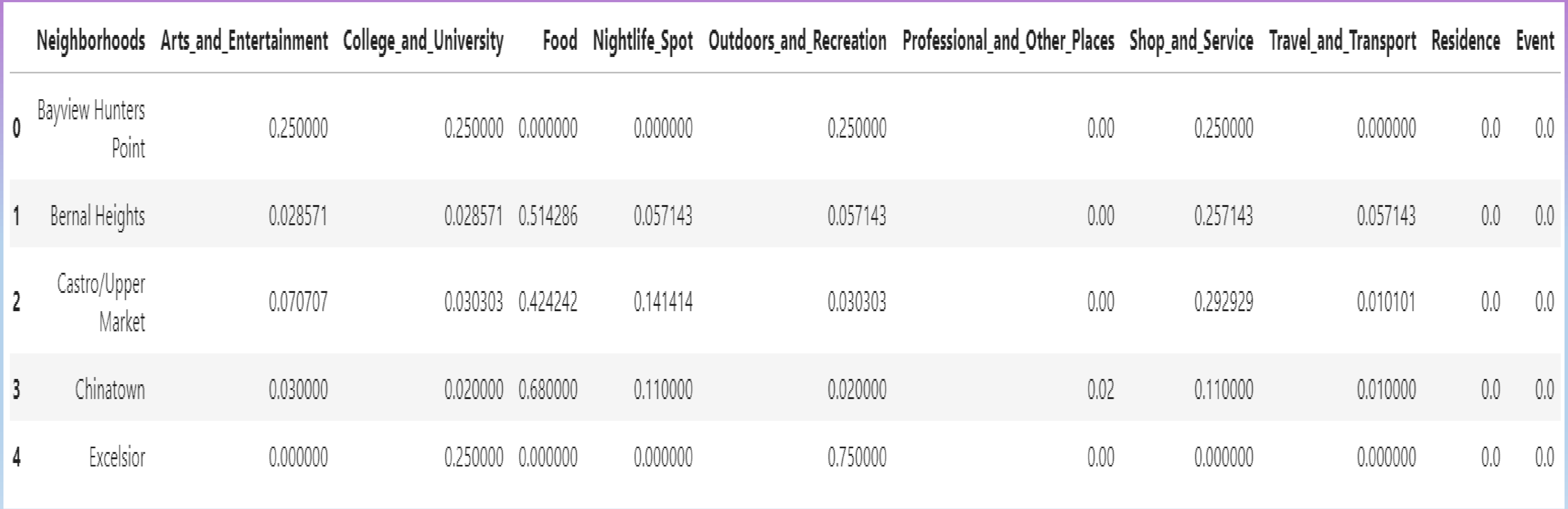
average (37)



outliers (6)

average (72)

Using percentage of primary categories of popular venues of neighborhoods (from Foursquare API) as training metric



	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

Balanced imbalanced classes with over-sampling

- **Naive**:

Increases the number of instances in the minority class (outliers) by replicating them in order to present a higher representation of the minority class in the sample.

- **Synthetic minority over-sampling technique, SMOTE**:

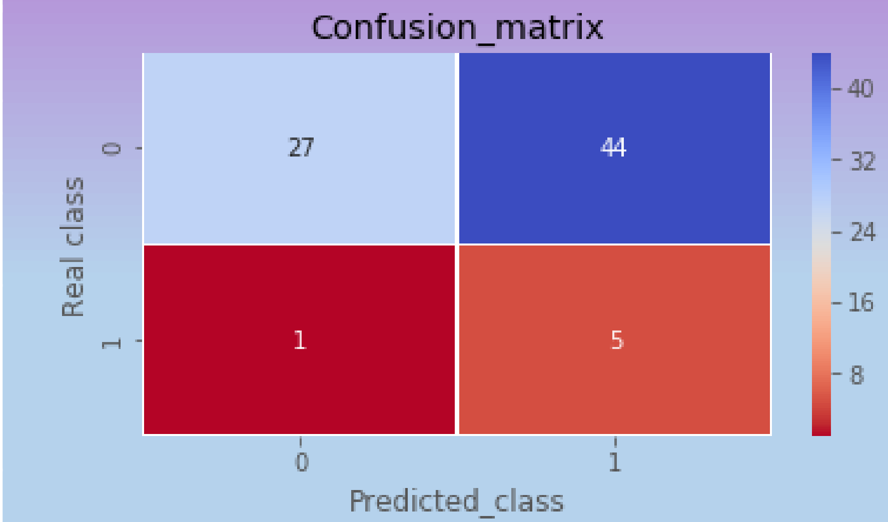
Takes a subset of data from the minority class as an example and creates new synthetic similar instances.

Desirable recall came with low precision of outliers

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

Best model

TN:FN≈1:9



Possible reasons responsible for the low precision of the model

- **Indistinct definition of credit card fraud:**

The criminal description of Chicago data is not very informative for credit card fraud, as the description was just 'credit card fraud'.

- **Small number of training data:**

Small number of training data enhance the dilemma of overfitting and underfitting.

Future direction

- 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.

- 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.