Crime rate analysis

1.Introduction

This small project will focus on analysis crime data set based on a public dataset(UCI Machine Learning Repository). In this data set it intend to predict number of crime per 100K population. Normally prediction is based on this single dataset, with foursquare API we can introduce more feature related to location data. With this help we can provide more accurate prediction. Although the dataset is very old, we still can see the way of handling these data. This method can also be used for new dataset, if it avalible publicly.

1.1Business understanding

The prediction can be widely used in different area such as estate retailer, governor, police management team and etc. Estate retailer can use this prediction as a reference in their advertisement. Also can use this number to set the price. Governor can decide how they plan to develop the city. Police department can reference this number to arrange patrol, where to setup new police station, how they assign police officer.

2.Data

2.1 Data Source

In this project I will use two dataset the first one is "Communities and Crime Data Set". This data is combined socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. It contains 1994 instances and 128 attributes, and it also contains missing value.
The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in incorrect values for per capita violent crime. These cities are not included in the dataset. Many of these omitted communities were from the midwestern USA. Data is described below based on original values. All numeric data was normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method.
Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small). E.g. An attribute described as 'mean people per household' is actually the normalized (0-1) version of that value. The normalization preserves rough ratios of values WITHIN an attribute (e.g. double the value for double the population within the available precision - except for extreme values (all values more than 3 SD above the mean are normalized to 1.00; all values more than 3 SD below the mean are normalized to 0.00)). However, the normalization does not preserve relationships between values BETWEEN attributes (e.g. it would not be meaningful to compare the value for whitePerCap with the value for blackPerCap for a community)
A limitation was that the LEMAS survey was of the police departments with at least 100 officers, plus a random sample of smaller departments. For our purposes, communities not found in both census and crime datasets were omitted.
Many communities are missing LEMAS data. Foursquare API will be used based on state county and community value. Venues will be retrieve for that community and grouped by number of place of interest, night club and hotel.

2.2 Data cleaning

This data set has two version. The one mainly used in this project is normalized one. But its location data is also masked, so I introduced the unnormalized data to get location. In unnormalized data use 'communityname' and 'state' to construct new location attribute 'address'. Then merge this attribute back to normalized data set. After merge, normalized data set it contains lots of missing values. 23 out of 126 attributes has missing value. Since all data are independent so I removed column which contains missing values. As a result, data become 1994 rows and 103 column including target.

2.3 Feature engineer

After cleaned data, I try to exam correlation among attribute and target value(total number of violent crimes per 100K population (numeric - decimal) GOAL attribute). I selected all attributes which correlation greater than 0.5 as input attribute. Now attribute left are 27 attributes. Then I use scatter diagram to view the relationship between input attribute with our target value. Diagrams can be viewed in my notebook. All feature selected looks correlated with targets.

27 attributes selected are
['PctKids2Par', 'racePctWhite', 'PctIlleg', 'NumIlleg', 'PctPopUnderPov', 'PctPersDenseHous', 'HousVacant', 'FemalePctDiv', 'pctWInvInc', 'agePct12t29', 'PctLargHouseOccup', 'racepctblack', 'OwnOccMedVal', 'PctFam2Par', 'PctUsePubTrans', 'medIncome', 'NumStreet', 'MalePctDivorce', 'PctImmigRec8', 'PctLargHouseFam', 'LandArea', 'PctNotHSGrad', 'pctWFarmSelf', 'perCapInc', 'agePct12t21', 'PctHousLess3BR', 'pctWPubAsst']

Then I try to build an regression model based on all these features. Data set is separated into train and test data set, and proportion is 8:2. Then apply xgboost regression model on data set. Use scikit-learn grid parameter search to find best parameter combination. Model on train set shows 0.63 R-square, but only have 0.32 on testing set. With this result model shows over-fit on training data set.

The problem may cause by incorrect feature selection. Try to use another way of feature selection. Train a XGBoost model on whole data set. And use predictive importance of this model to filter out features. In experiment, I tried importance greater than 0.04, 0.05, 0.06, 0.07 and 0.08. And selection five set of different features.

3. Data analysis

3.1 Relationship between target and percentage of population that is African American, percentage of population that is Caucasian

Scatter chart shows less percentage of African American will have lower crime rate. But these two didn't have obvious linear relationship. When percentage of African American lower than 0.2, most Target value is smaller than 0.4. The contrary the more percentage of population that is Caucasian, the lower crime rate is. When percentage of population that is Caucasian is greater than 0.8 crime rate is lower than 0.2.

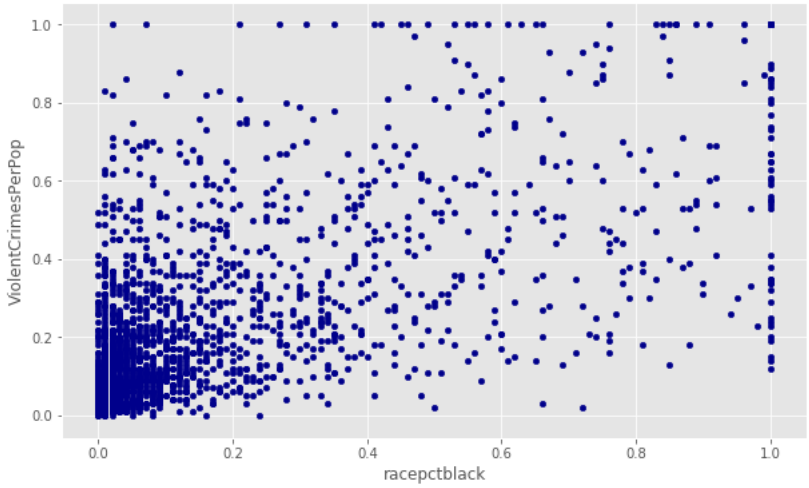

Figure 1 target and percentage of population that is African American

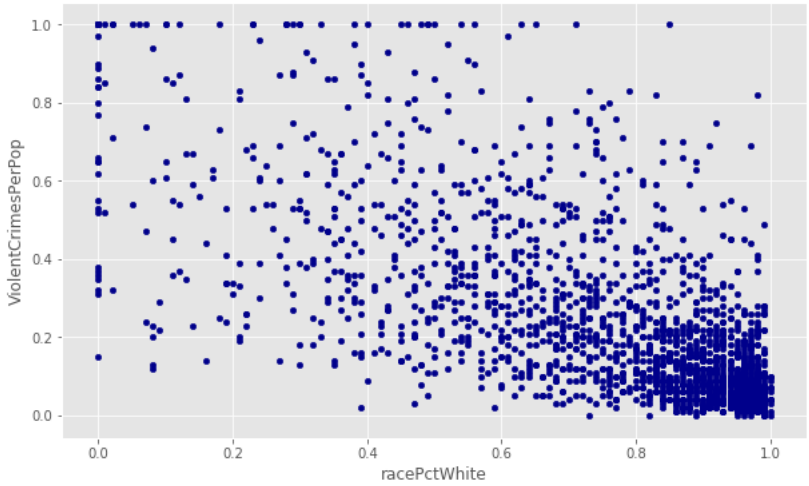

Figure 2 target and percentage of population that is Caucasian

3.2 Relationship between target and percentage of people under the poverty level, percentage of people 16 and over, in the labor force, and unemployed

Scatter chart shows how economy factors influence are target value. Less percentage of people under the poverty level will have lower crime rate and the same happens on unemployed rate. Which means when we want to control crime rate, we need to lower unemployed rate and increase people's income.

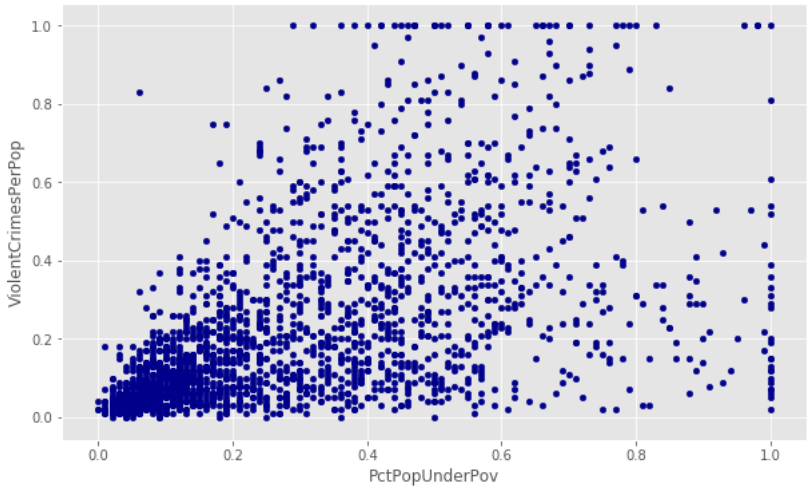

Figure 3 target and percentage of people under the poverty level

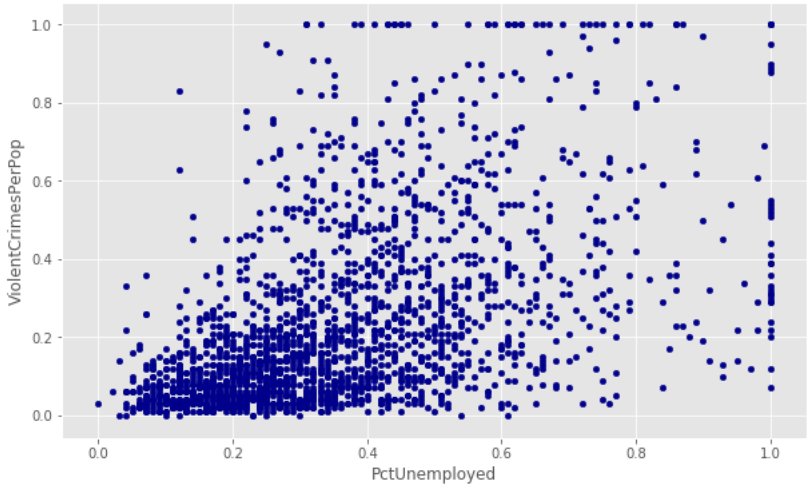

Figure 4 target and percentage of people 16 and over, in the labor force, and unemployed

3.3 Relationship between target and number of kids born to never married, number of vacant households

There are several interesting relationships in data following two are example. number of kids born to never married and number of vacant households are all very close to 0 (data is normalized), and these attribute also contribute to crime rate with high importance (according to).

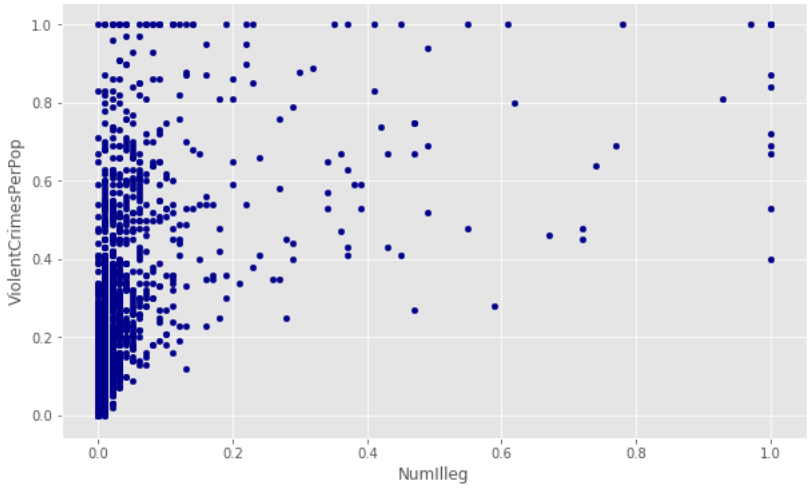

Figure 5 target and number of kids born to never married

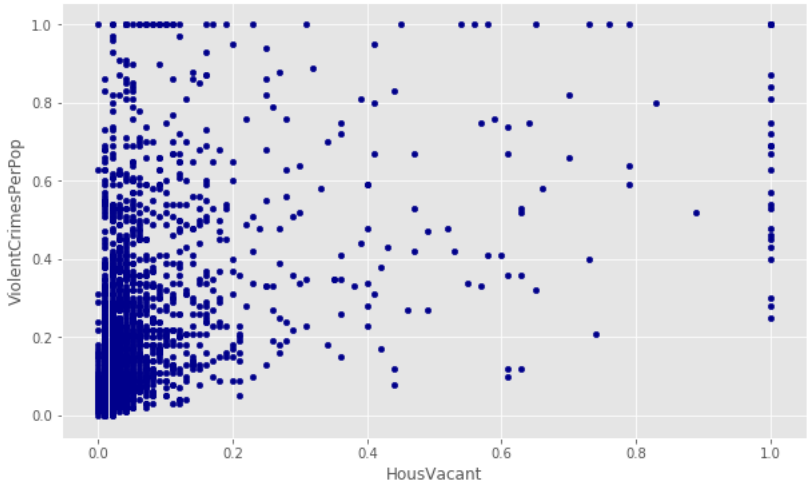

Figure 6 target and number of vacant households

4. Predictive Modeling

In this case our target is to predict crime rate pre-population, so regression model can fit this use case. Among different regression model I choose XGBoost (eXtreme Gradient Boosting) regression model. XGBoost is widely used for Kaggle competitions. The reason are easy to use, user can install it use pip command also it has been harvested in scikit-learn; have good efficiency on training; accuracy is good on different data set; user can easily tune parameters and get good result for different problems.

4.1 Tune XGBoost model

In feature engineer section, I have mention we also use another XGBoost model to help with select different feature for final model. Way of training this model is use all features in data set. Since parameter of XGBoost has lots of different combination, and its parameter will highly impact on model accuracy. So I use grid search function of scikit-learn package to search best parameter. I introduced 120 different combination in model training. {'max_depth': 2, 'n_estimators': 100} is the combination which has highest R2-score for both training and testing data set.

Then based on this best model I selected features with importance value above 0.006 as final model input feature. I also tried features with importance value above 0.004, 0.006, 0.007, 0.008. 0.005 has the best result.

For the final model it contains 22 features:

['PctKids2Par', 'PctIlleg', 'racePctWhite', 'NumImmig', 'PctPersDenseHous', 'FemalePctDiv', 'NumIlleg', 'PctPopUnderPov', 'HousVacant', 'PctLargHouseFam', 'pctWInvInc', 'medIncome', 'PctImmigRec5', 'MalePctDivorce', 'racepctblack', 'TotalPctDiv', 'numbUrban', 'PctBornSameState', 'RentMedian', 'PctImmigRec10', 'PctFam2Par', 'PctHousOccup']

And best parameter combination for final model is {'max_depth': 4, 'n_estimators': 50}, and R2-score on testing data set is 0.51.

5.Conclusion

In this study, I analyzed crime data of US. From final model result, model is not good enough, but from feature selected perspective we do have some finding on what impact on crime rate is. The key factors are from family problem, economic problem. So it is important to make sure helping solving people's family and economic problems.