# Title: Predictive Model: Crime Prediction.

**Type of Your Projects:** Data Analysis/Mining, Application Implement/Development, Research Project

1. **Introduction**

When it comes to security management, the authorities are always stretched for resources, as resources are scare and the scope of the responsibilities is very wide. Thus, resource allocation becomes an arduous task for the police department.

With the innovation of artificial intelligence and machine learning the police department can make use of AI to predict crime and use this predictive model to improve their resource allocation to the areas which are more prone to crime.

With this in mind we are trying to develop a model which will be able to predict crime rate based on economical and demographical information. We are planning to use following algorithms:

1. Decision Tree Classifier
2. Gaussian Naïve Bayes Classifier
3. Linear Support Vector Machine/Classifier
4. Linear Regression
5. Logistic Regression

Classification is a technique used to predict group membership for data instances. It is the task of generalizing known structure to apply to new data. As we have learned above algorithms in our Data Mining & Data Analytics classes, we wanted to implement them to predict crime rate & its factors.

1. **Data Sets**

Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, and so forth.

of

The data is extracted from UCI.com. The data is an amalgamation of social-economic data from 1990 US census, law enforcement data from 1990 US LEMAS survey and crime data from 1995 FBI UCR.

Varied attributes are included in the so that the model can select and learn from attributes, with this in mind complete unrelated attributes are ignored, only attributes which demonstrate plausible connection to crime are picked. These make up for the independent variables. Along with these dependent variables which are to be predicted are handpicked. The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units.

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.

|  |  |
| --- | --- |
| **Attribute Information** | |
|  |  |
| Attribute Information | (122 predictive, 5 non-predictive, 1 goal) |
| state | US state (by number) - not counted as predictive above, but if considered, should be consided nominal (nominal) |
| county | numeric code for county - not predictive, and many missing values (numeric) |
| community | numeric code for community - not predictive and many missing values (numeric) |
| communityname | community name - not predictive - for information only (string) |
| fold | fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric) |
| population | population for community |
| householdsize | mean people per household (numeric - decimal) |
| racepctblack | percentage of population that is african american (numeric - decimal) |
| racePctWhite | percentage of population that is caucasian (numeric - decimal) |
| racePctAsian | percentage of population that is of asian heritage (numeric - decimal) |
| racePctHisp | percentage of population that is of hispanic heritage (numeric - decimal) |
| agePct12t21 | percentage of population that is 12-21 in age (numeric - decimal) |
| agePct12t29 | percentage of population that is 12-29 in age (numeric - decimal) |
| agePct16t24 | percentage of population that is 16-24 in age (numeric - decimal) |
| agePct65up | percentage of population that is 65 and over in age (numeric - decimal) |
| numbUrban | number of people living in areas classified as urban (numeric - decimal) |
| pctUrban | percentage of people living in areas classified as urban (numeric - decimal) |
| medIncome | median household income (numeric - decimal) |
| pctWWage | percentage of households with wage or salary income in 1989 (numeric - decimal) |
| pctWFarmSelf | percentage of households with farm or self employment income in 1989 (numeric - decimal) |
| pctWInvInc | percentage of households with investment / rent income in 1989 (numeric - decimal) |
| pctWSocSec | percentage of households with social security income in 1989 (numeric - decimal) |
| pctWPubAsst | percentage of households with public assistance income in 1989 (numeric - decimal) |
| pctWRetire | percentage of households with retirement income in 1989 (numeric - decimal) |
| medFamInc | median family income (differs from household income for non-family households) (numeric - decimal) |
| perCapInc | per capita income (numeric - decimal) |
| whitePerCap | per capita income for caucasians (numeric - decimal) |
| blackPerCap | per capita income for african americans (numeric - decimal) |
| indianPerCap | per capita income for native americans (numeric - decimal) |
| AsianPerCap | per capita income for people with asian heritage (numeric - decimal) |
| OtherPerCap | per capita income for people with 'other' heritage (numeric - decimal) |
| HispPerCap | per capita income for people with hispanic heritage (numeric - decimal) |
| NumUnderPov | number of people under the poverty level (numeric - decimal) |
| PctPopUnderPov | percentage of people under the poverty level (numeric - decimal) |
| PctLess9thGrade | percentage of people 25 and over with less than a 9th grade education (numeric - decimal) |
| PctNotHSGrad | percentage of people 25 and over that are not high school graduates (numeric - decimal) |
| PctBSorMore | percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal) |
| PctUnemployed | percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal) |
| PctEmploy | percentage of people 16 and over who are employed (numeric - decimal) |
| PctEmplManu | percentage of people 16 and over who are employed in manufacturing (numeric - decimal) |
| PctEmplProfServ | percentage of people 16 and over who are employed in professional services (numeric - decimal) |
| PctOccupManu | percentage of people 16 and over who are employed in manufacturing (numeric - decimal) ######## |
| PctOccupMgmtProf | percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal) |
| MalePctDivorce | percentage of males who are divorced (numeric - decimal) |
| MalePctNevMarr | percentage of males who have never married (numeric - decimal) |
| FemalePctDiv | percentage of females who are divorced (numeric - decimal) |
| TotalPctDiv | percentage of population who are divorced (numeric - decimal) |
| PersPerFam | mean number of people per family (numeric - decimal) |
| PctFam2Par | percentage of families (with kids) that are headed by two parents (numeric - decimal) |
| PctKids2Par | percentage of kids in family housing with two parents (numeric - decimal) |
| PctYoungKids2Par | percent of kids 4 and under in two parent households (numeric - decimal) |
| PctTeen2Par | percent of kids age 12-17 in two parent households (numeric - decimal) |
| PctWorkMomYoungKids | percentage of moms of kids 6 and under in labor force (numeric - decimal) |
| PctWorkMom | percentage of moms of kids under 18 in labor force (numeric - decimal) |
| NumIlleg | number of kids born to never married (numeric - decimal) |
| PctIlleg | percentage of kids born to never married (numeric - decimal) |
| NumImmig | total number of people known to be foreign born (numeric - decimal) |
| PctImmigRecent | percentage of \_immigrants\_ who immigated within last 3 years (numeric - decimal) |
| PctImmigRec5 | percentage of \_immigrants\_ who immigated within last 5 years (numeric - decimal) |
| PctImmigRec8 | percentage of \_immigrants\_ who immigated within last 8 years (numeric - decimal) |
| PctImmigRec10 | percentage of \_immigrants\_ who immigated within last 10 years (numeric - decimal) |
| PctRecentImmig | percent of \_population\_ who have immigrated within the last 3 years (numeric - decimal) |
| PctRecImmig5 | percent of \_population\_ who have immigrated within the last 5 years (numeric - decimal) |
| PctRecImmig8 | percent of \_population\_ who have immigrated within the last 8 years (numeric - decimal) |
| PctRecImmig10 | percent of \_population\_ who have immigrated within the last 10 years (numeric - decimal) |
| PctSpeakEnglOnly | percent of people who speak only English (numeric - decimal) |
| PctNotSpeakEnglWell | percent of people who do not speak English well (numeric - decimal) |
| PctLargHouseFam | percent of family households that are large (6 or more) (numeric - decimal) |
| PctLargHouseOccup | percent of all occupied households that are large (6 or more people) (numeric - decimal) |
| PersPerOccupHous | mean persons per household (numeric - decimal) |
| PersPerOwnOccHous | mean persons per owner occupied household (numeric - decimal) |
| PersPerRentOccHous | mean persons per rental household (numeric - decimal) |
| PctPersOwnOccup | percent of people in owner occupied households (numeric - decimal) |
| PctPersDenseHous | percent of persons in dense housing (more than 1 person per room) (numeric - decimal) |
| PctHousLess3BR | percent of housing units with less than 3 bedrooms (numeric - decimal) |
| MedNumBR | median number of bedrooms (numeric - decimal) |
| HousVacant | number of vacant households (numeric - decimal) |
| PctHousOccup | percent of housing occupied (numeric - decimal) |
| PctHousOwnOcc | percent of households owner occupied (numeric - decimal) |
| PctVacantBoarded | percent of vacant housing that is boarded up (numeric - decimal) |
| PctVacMore6Mos | percent of vacant housing that has been vacant more than 6 months (numeric - decimal) |
| MedYrHousBuilt | median year housing units built (numeric - decimal) |
| PctHousNoPhone | percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal) |
| PctWOFullPlumb | percent of housing without complete plumbing facilities (numeric - decimal) |
| OwnOccLowQuart | owner occupied housing - lower quartile value (numeric - decimal) |
| OwnOccMedVal | owner occupied housing - median value (numeric - decimal) |
| OwnOccHiQuart | owner occupied housing - upper quartile value (numeric - decimal) |
| RentLowQ | rental housing - lower quartile rent (numeric - decimal) |
| RentMedian | rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal) |
| RentHighQ | rental housing - upper quartile rent (numeric - decimal) |
| MedRent | median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal) |
| MedRentPctHousInc | median gross rent as a percentage of household income (numeric - decimal) |
| MedOwnCostPctInc | median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal) |
| MedOwnCostPctIncNoMtg | median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal) |
| NumInShelters | number of people in homeless shelters (numeric - decimal) |
| NumStreet | number of homeless people counted in the street (numeric - decimal) |
| PctForeignBorn | percent of people foreign born (numeric - decimal) |
| PctBornSameState | percent of people born in the same state as currently living (numeric - decimal) |
| PctSameHouse85 | percent of people living in the same house as in 1985 (5 years before) (numeric - decimal) |
| PctSameCity85 | percent of people living in the same city as in 1985 (5 years before) (numeric - decimal) |
| PctSameState85 | percent of people living in the same state as in 1985 (5 years before) (numeric - decimal) |
| LemasSwornFT | number of sworn full time police officers (numeric - decimal) |
| LemasSwFTPerPop | sworn full time police officers per 100K population (numeric - decimal) |
| LemasSwFTFieldOps | number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - decimal) |
| LemasSwFTFieldPerPop | sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal) |
| LemasTotalReq | total requests for police (numeric - decimal) |
| LemasTotReqPerPop | total requests for police per 100K popuation (numeric - decimal) |
| PolicReqPerOffic | total requests for police per police officer (numeric - decimal) |
| PolicPerPop | police officers per 100K population (numeric - decimal) |
| RacialMatchCommPol | a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal) |
| PctPolicWhite | percent of police that are caucasian (numeric - decimal) |
| PctPolicBlack | percent of police that are african american (numeric - decimal) |
| PctPolicHisp | percent of police that are hispanic (numeric - decimal) |
| PctPolicAsian | percent of police that are asian (numeric - decimal) |
| PctPolicMinor | percent of police that are minority of any kind (numeric - decimal) |
| OfficAssgnDrugUnits | number of officers assigned to special drug units (numeric - decimal) |
| NumKindsDrugsSeiz | number of different kinds of drugs seized (numeric - decimal) |
| PolicAveOTWorked | police average overtime worked (numeric - decimal) |
| LandArea | land area in square miles (numeric - decimal) |
| PopDens | population density in persons per square mile (numeric - decimal) |
| PctUsePubTrans | percent of people using public transit for commuting (numeric - decimal) |
| PolicCars | number of police cars (numeric - decimal) |
| PolicOperBudg | police operating budget (numeric - decimal) |
| LemasPctPolicOnPatr | percent of sworn full time police officers on patrol (numeric - decimal) |
| LemasGangUnitDeploy | gang unit deployed (numeric - decimal - but really ordinal - 0 means NO, 1 means YES, 0.5 means Part Time) |
| LemasPctOfficDrugUn | percent of officers assigned to drug units (numeric - decimal) |
| PolicBudgPerPop | police operating budget per population (numeric - decimal) |
| ViolentCrimesPerPop | total number of violent crimes per 100K popuation (numeric - decimal) GOAL attribute (to be predicted) |

1. **Research Problems**

* We will first need to classify the data which willcheck if there is high crime rate in a given locality and check what is the proportion of high crime rate and low crime rate.
* We are trying to find out top 10 features which are responsible for crime.
* We want to compare the models depending on their outputs.
* Another issue could be overfitting of the model. Since the data is huge, we need to find out how much data we should be taking for training, testing & validation. When we will be testing the model, there is possibility that it might fail due to overfitting problem.

1. **Potential Solutions**

To find the best predictive features we will be using Decision Tree Classifier, Naïve Bayes, Linear Support Vector Machine, Linear Regression, Logistic Regression.

And we will compare the accuracy the results of each model with one another, to do so we will analyze the accuracy, precision and recall of the model before and after cross validation.

Depending on the classification algorithm, we will try to predict top 10 features for every algorithm. Then we are planning to compare output of the algorithms depending on their precision, accuracy & recall. After comparison, we will decide whichmodel performed better in terms of precision & accuracy.

* When using Naïve Bayes classifier to predict high crime we will decide top predictive features by measuring the normalized absolute difference of mean of the features between two classes.
* When using Support Vector Machine model to predict high crime we will decide top predictive features by measuring the absolute feature weight in the model.
* When using Linear Regression model to predict high crime we will measure the MSE of the model.
* We will use Ridge Regression to reduce the amount of overfitting.

1. **Evaluations**

We are planning to use cross validations of 10 folds to evaluate accuracy, precision & recall.

Depending on their outputs, we will consider mean accuracy, mean precision & mean recall. It will compare it with other models to determine the performance. Also, we can determine which are common features in the model which contribute to the crimes. We will evaluate different classification model based on their accuracy of the model. And how do they compare.

1. **Expected Outcomes**

The model will predict the crime rate & features which contribute more to the crime areas based on following algorithms:

1. Decision Tree Classifier
2. Gaussian Naïve Bayes Classifier
3. Linear Support Vector Machine/Classifier
4. Linear Regression
5. Logistic Regression

We will be comparing their outputs to find out which model provides us a better accuracy & precision.