Crime Prediction Model Report

Prepared by: Alperen Unal

1. Introduction

- The goal of this analysis is to develop a predictive model for per capita violent crimes in communities based on various attributes such as urban population and police statistics. The objective is to create an interpretable model that can inform crime prevention strategies. The analysis involves the exploration of the dataset, data cleaning, feature selection, and the creation of predictive models using linear regression and k-nearest neighbors (KNN).

- The project aims to predict per capita violent crimes in communities based on various attributes, focusing on building an interpretable model to inform crime prevention strategies.

2. Dataset Overview

The dataset is from the UCI Machine Learning Repository, specifically the "Communities and Crime" dataset.

Dataset link: https://archive.ics.uci.edu/dataset/183/communities+and+crime
Dataset creator: Michael Redmond

- 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 these purposes, communities not found in both census and crime datasets were omitted. Many communities are missing LEMAS data.

The dataset contains information on different communities, including demographic, economic, and law enforcement-related attributes. The attributes have been anonymized for privacy reasons.

2.1. Information about dataset's variables:

Non-predictive attributes:

- **state**: US state (by number) - nominal (nominal)
- **county**: numeric code for county - numeric
- **community**: numeric code for community - numeric
- **communityname**: community name - string
- **fold**: fold number for non-random 10-fold cross-validation – numeric

Predictive attributes:

- **population**: population for community - decimal numeric
- **householdsize**: mean people per household - decimal numeric
- **racepctblack**: percentage of population that is African American - decimal numeric
- **racePctWhite**: percentage of population that is Caucasian - decimal numeric
- **racePctAsian**: percentage of population that is of Asian heritage - decimal numeric
- **racePctHisp**: percentage of population that is of Hispanic heritage - decimal numeric
- **agePct12t21**: percentage of population that is 12-21 in age - decimal numeric
- **agePct12t29**: percentage of population that is 12-29 in age - decimal numeric

- **agePct16t24**: percentage of population that is 16-24 in age - decimal numeric
- **agePct65up**: percentage of population that is 65 and over in age - decimal numeric
- **numbUrban**: number of people living in areas classified as urban - decimal numeric
- **pctUrban**: percentage of people living in areas classified as urban - decimal numeric
- **medIncome**: median household income - decimal numeric
- **pctWWage**: percentage of households with wage or salary income in 1989 - decimal numeric
- **pctWFarmSelf**: percentage of households with farm or self-employment income in 1989 - decimal numeric
- **pctWInvInc**: percentage of households with investment/rent income in 1989 - decimal numeric
- **pctWSocSec**: percentage of households with social security income in 1989 - decimal numeric
- **pctWPubAsst**: percentage of households with public assistance income in 1989 - decimal numeric
- **pctWRetire**: percentage of households with retirement income in 1989 - decimal numeric
- **medFamInc**: median family income (differs from household income for non-family households) - decimal numeric
- **perCapInc**: per capita income - decimal numeric
- **whitePerCap**: per capita income for Caucasians - decimal numeric
- **blackPerCap**: per capita income for African Americans - decimal numeric
- **indianPerCap**: per capita income for Native Americans - decimal numeric
- **AsianPerCap**: per capita income for people with Asian heritage - decimal numeric
- **OtherPerCap**: per capita income for people with 'other' heritage - decimal numeric
- **HispPerCap**: per capita income for people with Hispanic heritage - decimal numeric
- **NumUnderPov**: number of people under the poverty level - decimal numeric
- **PctPopUnderPov**: percentage of people under the poverty level - decimal numeric
- **PctLess9thGrade**: percentage of people 25 and over with less than a 9th-grade education - decimal numeric
- **PctNotHSGrad**: percentage of people 25 and over that are not high school graduates - decimal numeric
- **PctBSorMore**: percentage of people 25 and over with a bachelor's degree or higher education - decimal numeric
- **PctUnemployed**: percentage of people 16 and over, in the labor force, and unemployed - decimal numeric
- **PctEmploy**: percentage of people 16 and over who are employed - decimal numeric
- **PctEmplManu**: percentage of people 16 and over who are employed in manufacturing - decimal numeric
- **PctEmplProfServ**: percentage of people 16 and over who are employed in professional services - decimal numeric
- **PctOccupManu**: percentage of people 16 and over who are employed in manufacturing - decimal numeric
- **PctOccupMgmtProf**: percentage of people 16 and over who are employed in management or professional occupations - decimal numeric
- **MalePctDivorce**: percentage of males who are divorced - decimal numeric
- **MalePctNevMarr**: percentage of males who have never married - decimal numeric
- **FemalePctDiv**: percentage of females who are divorced - decimal numeric
- **TotalPctDiv**: percentage of the population who are divorced - decimal numeric
- **PersPerFam**: mean number of people per family - decimal numeric
- **PctFam2Par**: percentage of families (with kids) that are headed by two parents - decimal numeric
- **PctKids2Par**: percentage of kids in family housing with two parents - decimal numeric
- **PctYoungKids2Par**: percentage of kids 4 and under in two-parent households - decimal numeric
- **PctTeen2Par**: percentage of kids age 12-17 in two-parent households - decimal numeric
- **PctWorkMomYoungKids**: percentage of moms of kids 6 and under in the labor force - decimal numeric
- **PctWorkMom**: percentage of moms of kids under 18 in the labor force - decimal numeric
- **NumIlleg**: number of kids born to never married - decimal numeric
- **PctIlleg**: percentage of kids born to never married - decimal numeric
- **NumImmig**: total number of people known to be foreign-born - decimal numeric

- **PctImmigRecent**: percentage of immigrants who immigrated within the last 3 years - decimal numeric
- **PctImmigRec5**: percentage of immigrants who immigrated within the last 5 years - decimal numeric
- **PctImmigRec8**: percentage of immigrants who immigrated within the last 8 years - decimal numeric
- **PctImmigRec10**: percentage of immigrants who immigrated within the last 10 years - decimal numeric
- **PctRecentImmig**: percent of the population who have immigrated within the last 3 years - decimal numeric
- **PctRecImmig5**: percent of the population who have immigrated within the last 5 years - decimal numeric
- **PctRecImmig8**: percent of the population who have immigrated within the last 8 years - decimal numeric
- **PctRecImmig10**: percent of the population who have immigrated within the last 10 years - decimal numeric
- **PctSpeakEnglOnly**: percent of people who speak only English - decimal numeric
- **PctNotSpeakEnglWell**: percent of people who do not speak English well - decimal numeric
- **PctLargHouseFam**: percent of family households that are large (6 or more) - decimal numeric
- **PctLargHouseOccup**: percent of all occupied households that are large (6 or more people) - decimal numeric
- **PersPerOccupHous**: mean persons per household - decimal numeric
- **PersPerOwnOccHous**: mean persons per owner-occupied household - decimal numeric
- **PersPerRentOccHous**: mean persons per rental household - decimal numeric
- **PctPersOwnOccup**: percent of people in owner-occupied households - decimal numeric
- **PctPersDenseHous**: percent of persons in dense housing (more than 1 person per room) - decimal numeric
- **PctHousLess3BR**: percent of housing units with less than 3 bedrooms - decimal numeric
- **MedNumBR**: median number of bedrooms - decimal numeric
- **HousVacant**: number of vacant households - decimal numeric
- **PctHousOccup**: percent of housing occupied - decimal numeric
- **PctHousOwnOcc**: percent of households owner-occupied - decimal numeric
- **PctVacantBoarded**: percent of vacant housing that is boarded up - decimal numeric
- **PctVacMore6Mos**: percent of vacant housing that has been vacant more than 6 months - decimal numeric
- **MedYrHousBuilt**: median year housing units built - decimal numeric
- **PctHousNoPhone**: percent of occupied housing units without phone (in 1990, this was rare!) - decimal numeric
- **PctWOFullPlumb**: percent of housing without complete plumbing facilities - decimal numeric
- **OwnOccLowQuart**: owner-occupied housing - lower quartile value - decimal numeric
- **OwnOccMedVal**: owner-occupied housing - median value - decimal numeric
- **OwnOccHiQuart**: owner-occupied housing - upper quartile value - decimal numeric
- **RentLowQ**: rental housing - lower quartile rent - decimal numeric
- **RentMedian**: rental housing - median rent (Census variable H32B from file STF1A) - decimal numeric
- **RentHighQ**: rental housing - upper quartile rent - decimal numeric
- **MedRent**: median gross rent (Census variable H43A from file STF3A - includes utilities) - decimal numeric
- **MedRentPctHousInc**: median gross rent as a percentage of household income - decimal numeric
- **MedOwnCostPctInc**: median owners' cost as a percentage of household income - for owners with a mortgage - decimal numeric
- **MedOwnCostPctIncNoMtg**: median owners' cost as a percentage of household income - for owners without a mortgage - decimal numeric
- **NumInShelters**: number of people in homeless shelters - decimal numeric
- **NumStreet**: number of homeless people counted in the street - decimal numeric
- **PctForeignBorn**: percent of people foreign-born - decimal numeric
- **PctBornSameState**: percent of people born in the same state as currently living - decimal numeric
- **PctSameHouse85**: percent of people living in the same house as in 1985 (5 years before) - decimal numeric
- **PctSameCity85**: percent of people living in the same city as in 1985 (5 years before) - decimal numeric

- **PctSameState85**: percent of people living in the same state as in 1985 (5 years before) - decimal numeric
- **LemasSwornFT**: number of sworn full-time police officers - decimal numeric
- **LemasSwFTPerPop**: sworn full-time police officers per 100K population - decimal numeric
- **LemasSwFTFieldOps**: number of sworn full-time police officers in field operations (on the street as opposed to administrative, etc.) - decimal numeric
- **LemasSwFTFieldPerPop**: sworn full-time police officers in field operations (on the street as opposed to administrative, etc.) per 100K population - decimal numeric
- **LemasTotalReq**: total requests for police - decimal numeric
- **LemasTotReqPerPop**: total requests for police per 100K population - decimal numeric
- **PolicReqPerOffic**: total requests for police per police officer - decimal numeric
- **PolicPerPop**: police officers per 100K population - decimal numeric
- **RacialMatchCommPol**: a measure of the racial match between the community and the police force. High values indicate proportions in the community and police force are similar - decimal numeric
- **PctPolicWhite**: percent of police that are Caucasian - decimal numeric
- **PctPolicBlack**: percent of police that are African American - decimal numeric
- **PctPolicHisp**: percent of police that are Hispanic - decimal numeric
- **PctPolicAsian**: percent of police that are Asian - decimal numeric
- **PctPolicMinor**: percent of police that are a minority of any kind - decimal numeric
- **OfficAssgnDrugUnits**: number of officers assigned to special drug units - decimal numeric
- **NumKindsDrugsSeiz**: number of different kinds of drugs seized - decimal numeric
- **PolicAveOTWorked**: police average overtime worked - decimal numeric
- **LandArea**: land area in square miles - decimal numeric
- **PopDens**: population density in persons per square mile - decimal numeric
- **PctUsePubTrans**: percent of people using public transit for commuting - decimal numeric
- **PolicCars**: number of police cars - decimal numeric
- **PolicOperBudg**: police operating budget - decimal numeric
- **LemasPctPolicOnPatr**: percent of sworn full-time police officers on patrol - decimal numeric
- **LemasGangUnitDeploy**: gang unit deployed - decimal numeric - but really ordinal - 0 means NO, 1 means YES, 0.5 means Part Time
- **LemasPctOfficDrugUn**: percent of officers assigned to drug units - decimal numeric
- **PolicBudgPerPop**: police operating budget per population - decimal numeric

Goal variable:

- **ViolentCrimesPerPop**: total number of violent crimes per 100K population - decimal numeric (GOAL attribute - to be predicted)

3. Data Pre-processing

3.1. Data Import and Initial Setup:

- The dataset was imported and assigned to **df**.
- The last two rows were removed to ensure data equality.

3.2. Data Splitting:

- The dataset was randomly split into training, validation, and test sets using a 1/3rd strategy.

3.3. Initial Data Analysis:

- The training set contained a specified number of records and variables.
- Input attributes and their types were identified and displayed.

3.4. Renaming Columns:

- Columns were renamed for better understanding and consistency across different sets.

3.5. Handling Missing Values:

- Missing values (represented as "?") were replaced with **NA**.
- Missing value percentages were calculated for each dataset.

4. Exploratory Data Analysis (EDA)

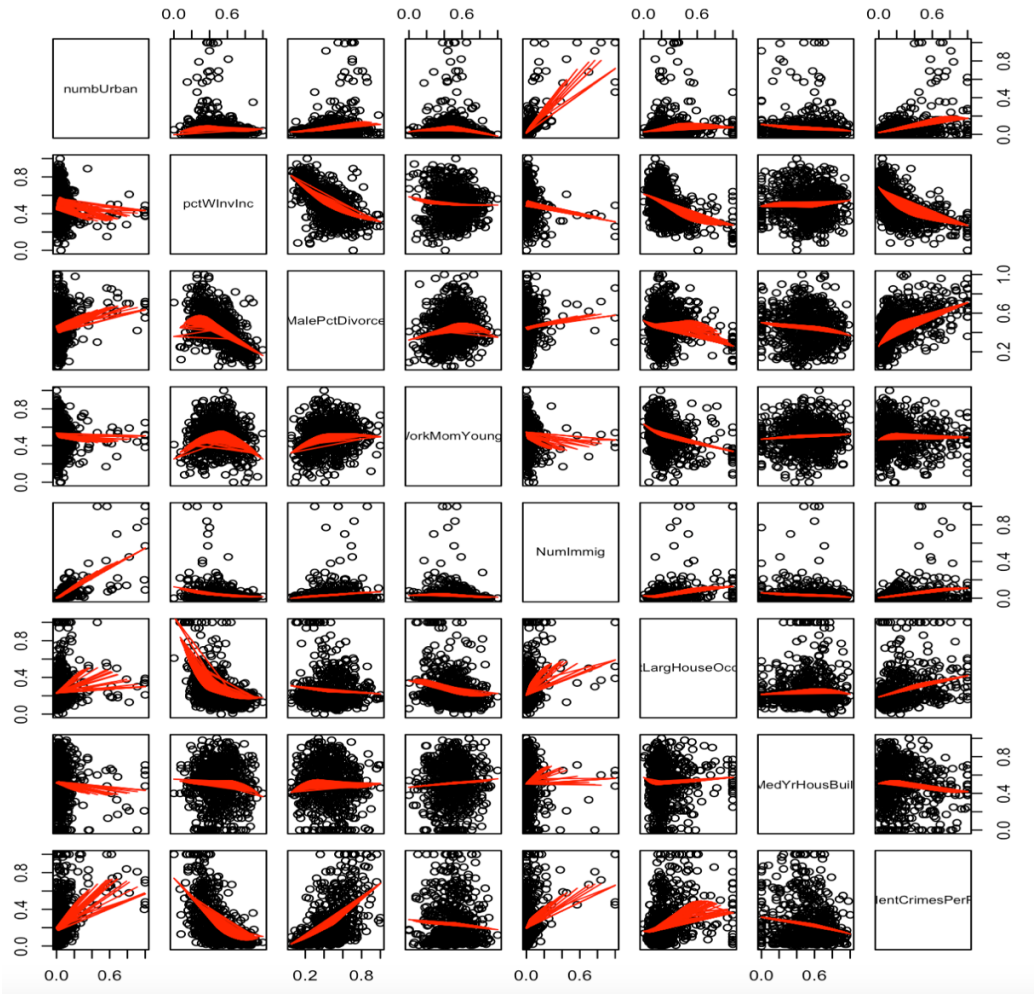
4.1. Correlation Analysis:

- Correlation matrix was created for predictive variables.
- Variables with the highest correlation with the target variable "ViolentCrimesPerPop" were identified.

4.2. Scatterplot Matrices:

- Scatterplot matrices were created for different sets of variables to visualize relationships.

Example:



5. Data Cleaning

5.1. Removal of Non-Predictive Attributes:

- Attributes like "county" and "community" were removed due to high missing values.

5.2. Discarding Low Correlation Attributes:

- Attributes without significant correlation with the target variable were identified and removed.

5.3. Missing Value Handling in Test Set:

- The missing value in the "OtherPerCap" column in the test set was replaced with its mean.

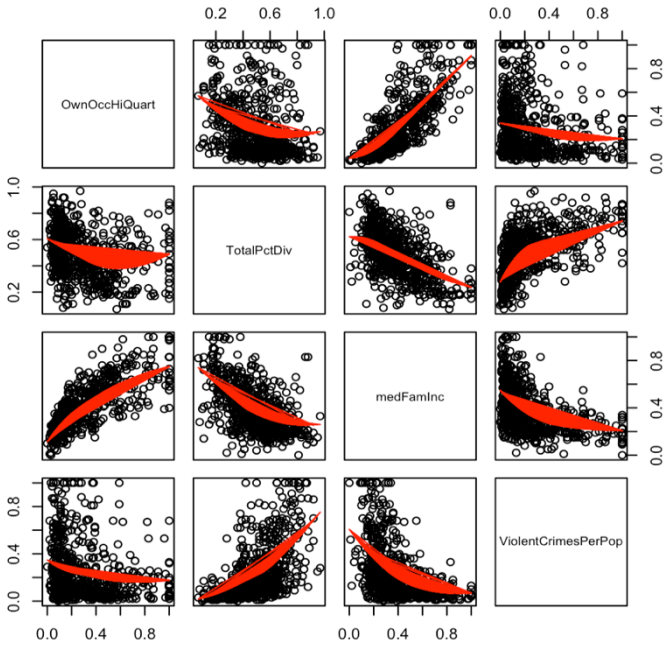
6. Data Cleaning

6.1. Linear Regression Model:

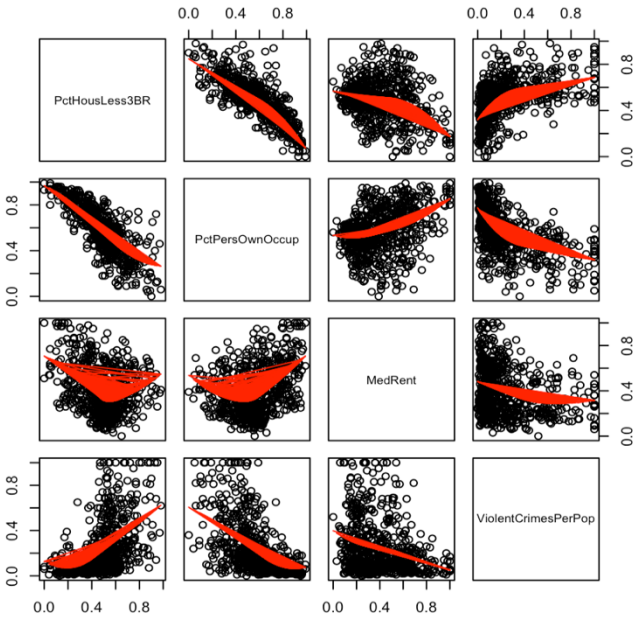
- The linear regression model was built and refined by removing variables with high p-values and low statistical significance.
- The final model (**Mb**) included all statistically significant predictors.

6.2. Polynomial Models:

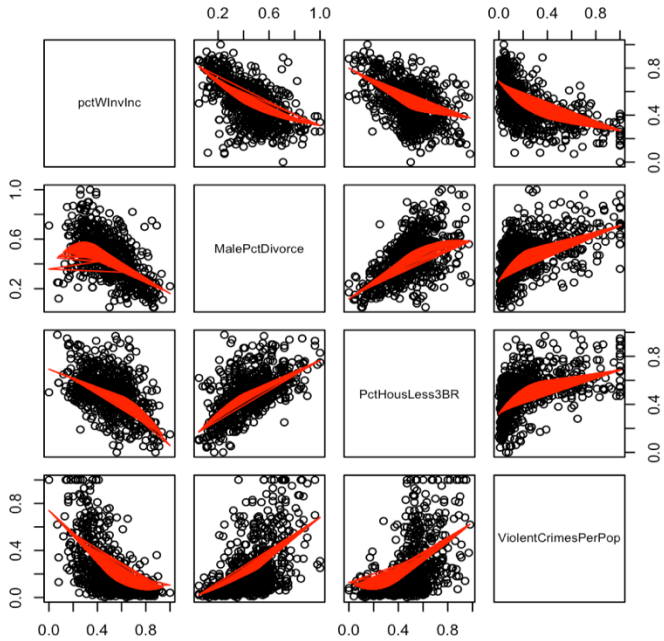
- Polynomial models were created by transforming selected predictors into polynomial features. This approach helps in capturing the non-linear relationships between predictors and the dependent variable.
- *Model M2:* Formed with second-degree polynomial terms of '**medFamInc'**, '**OwnOccHiQuart'**, and '**TotalPctDiv'**.



- *Model M3:* Consisted of second-degree polynomial terms of '**PctHousLess3BR'**, '**PctPersOwnOccup'**, and '**MedRent'**.



- *Model M4:* Included second-degree polynomial terms of '**pctWInvInc'**, '**MalePctDivorce'**, and '**PctHousLess3BR'**.



6.3. K-Nearest Neighbors (KNN) Models:

- KNN models were built for different values of k (1 to 4).
- The optimal value of k was determined based on the Mean Squared Error (MSE).

7. Model Evaluation:

7.1. Linear Regression (M1): The linear regression model (M1) has the lowest Mean Squared Error (MSE) on the validation set (0.0226), indicating it is the best predictor among the models tested.

7.2. Polynomial Models (M2, M3, M4):

M2: MSE = 0.03031253
This model includes second-degree polynomial terms of medFamInc, OwnOccHiQuart, and TotalPctDiv.
M3: MSE = 0.03697561
This model uses second-degree polynomial terms of PctHousLess3BR, PctPersOwnOccup, and MedRent.
M4: MSE = 0.02864897
This model includes second-degree polynomial terms of pctWInvInc, MalePctDivorce, and PctHousLess3BR.

7.3. K-Nearest Neighbors: The KNN model with k=1 provided the lowest MSE (0.0545) among the different KNN models, but it was still higher than the linear regression models.

8. Final Model Performance:

The final linear regression model (Mb) exhibited a Mean Squared Error (MSE) of 0.0212 on the test set, indicating a high level of accuracy.