Tucson Crimes and Wealth: Classification and Prediction Using Temporal Data, Regional Data, and Personal Traits

Trevor Nemetz and Nathan Crutchfield

University of Arizona
12/10/24

1 Introduction

Crime is a big issue in today's society, and it is a problem that can hit close to home, even here on the University of Arizona campus. Crime is something we want to be stopped of course, but doing so is no easy task. Given there are only a limited amount of police and community resources and an ever-greater pool of criminals, trying to stay ahead of the crime curve can be quite difficult. Given such substantial amounts of police report data has been generated over the years due to this fact, what if we could utilize that data to predict where crimes are more likely to occur? To pursue this, we would need to analyze which factors are most influential on crimes, and how those factors also influence the type of crime committed as well. Our goal for this project is to use those factors to find correlations between a person's age (AGE), sex of the person (SEX), the day of the week (DAY), the month (MONTH), the time of day (TIME), the wealth of a neighborhood (WEALTH), number of houses in a neighborhood (HOUSES), and the type of crime committed (TYPE). We want to use this crime data to discover patterns in their relationships that can allow us to predict beforehand what TYPE of crimes are more likely to be committed due to AGE, SEX, TIME, MONTH, DAY, and WEALTH, and utilize the DAY, MONTH, TIME, AGE, SEX, and HOUSES in an area where a crime is committed to predict the WEALTH. We assert that a dependency exists between these factors, and aim to verify, via training and testing machine learning models, if that is the case. If we can determine that such correlations exist and can predict relatively accurately, than we could be able to predetermine which areas are more crime-prone, and thus be able to more efficiently allocate crime-fighting and community resources.

2 Related Work

While we did not find any research that is the extremely close to what we intend to do, we did find supporting resources that helped us to shape our hypotheses and defined the path for the modeling. For example, in one study done by E. Britt Patterson in 1991, they "examined the relative importance of poverty and income inequality in explaining criminal activity across social areas. The results indicate that levels of absolute poverty, measured by the percentage of households with annual incomes below $5,000, are significantly associated with higher rates of serious violent crime" (769). These results show us that, as expected, low-income areas are more crime-prone. This conclusion helped us to decide that it would be much more beneficial to examine correlation between wealth with crime type instead of just redoing previous research. In doing so, we could build off their research and expand it instead of just validating. Additionally, we found a study that was conducted to determine he viability of machine learning in predicting and modeling criminal activities, which showed us that "machine learning models have proven to be effective for crime prediction", and the paper also touched on the success of models utilizing categorical regressions when predicting crimes and associated criminal data (Mandalapu et al.). With this in mind, we concluded that using logistic and linear regression models would be relevant when testing our hypotheses given the previous success of modeling crime as enumerated in their paper. Given that we did not exactly find research correlating these specific factors, and the fact that we only have access to data in Tucson, AZ, our research is a more targeted and expanded variation on Patterson's work, where we aim to use more factors to determine the type of crime likely to be committed. We also utilize similar systems of modeling that are mentioned in Mandalapu et al.'s writing, however the main novelty in that case is the targeting of the models' training for Tucson-specific crimes, which targets a different demographic of people, different living conditions, etc.... Overall, instead of focusing on just "will a crime be committed" or "will a crime of a specific type be committed", we intend to generalize the output and determine if the input criteria have an effect on the type of crime committed as a whole, predicting the type of crime instead of if a crime happens.

3 Methods

3.1 Background

3.1.1 Linear Models

A problem with a continuous output can be represented by a linear model. A linear model with n inputs has the form:

$$\hat{y} = \hat{w}_1 x_1 + \hat{w}_2 x_2 + ... + \hat{w}_n x_n \tag{1}$$

Each $\hat{w}_i$ is the coefficient for the input feature x_i, and helps interpret the importance of x_i compared to the other features. The weights can be found by minimizing the *loss* of the predicted and expected values. As referenced in a 2024 paper, "as in the case of outliers in the datasets, Mean Squared Error is the best strategy." (Aryan) Mean squared error is defined by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

The weights found after minimizing can then be interpreted. However, first the model needs to be evaluated, commonly based on its coefficient of determination, R^2.

$$R^2 = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

R^2 represents the proportion of the variability of the actual data (y) that is explained by the model ($\hat{y}$). Additionally, the residuals, $\vec{r}$, are useful in interpreting the data.

$$\vec{r} = y - \hat{y} = \begin{bmatrix} y_0 - \hat{y}_0 \\ y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \tag{4}$$

Additionally, L1 (Lasso) and L2 (Ridge) regularization can be used to restrict the model complexity to help prevent overfitting for linear models.

Given the relevance to the dataset, both inputs and outputs, a linear model is a relevant choice for the second hypothesis.

3.1.2 Logistic Models

A model with a positive and negative class can be represented by a logistic model. This follows the logistic curve, and the model is defined by:

$$\hat{p} = \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2 + ...)}} \tag{5}$$

$\hat{p}$ represents the probability of being the positive class given the input features. Since this is a classed based model, MSE is not a useful choice for the loss function. As mentioned in a 2020 paper, "the binary cross entropy is very convenient to train a model to solve many classification problems at the same time, if each classification can be reduced to a binary choice." (Ruby 5396) Cross entropy loss is defined by:

$$CE = -\sum_{i=1}^{n}(y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)) \tag{6}$$

However, since the first hypothesis depends on more than one input, one-vs-rest (OVR) logistic regression is used. OVR is defined by comparing each class against not being that class. Then, the class with the highest probability is chosen. Let n_o be the number of outputs ($n_o = 5$ for the *fel_misd* data), there will be n_o classifiers trained. Let $\hat{p}_0$ be the probability of the 0 class, $\hat{p}_1$ be the probability of the 1 class, ..., $\hat{p}_{n_o}$ be the probability of the n_o class, then the OVR classifier will find the class, K:

$$\max(\hat{p}_0, \hat{p}_1, ..., \hat{p}_{n_o}) = \hat{p}_i \implies K = i \tag{7}$$

Accuracy is a useful metric to evaluate the performance of a classifier, where accuracy is defined as:

$$\text{accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Number of Samples}} \tag{8}$$

Based on the relevant research and the data, a OVR logistic regression is a reasonable choice for this problem.

3.1.3 Cross Validation

K-fold cross validation is a useful tool for hyper-parameter optimization. Additionally, this helps to prevent overfitting, alognside finding the best performing model. The training data is split into k subsets and multiple models are trained with the different subset, and are evaluated with the resulting $k - 1$ subsets.

3.1.4 Correlation

Since $\hat{y}$ is dependent on multiple input features, x_i, it is important to guarantee that
$\text{Corr}(x_i, x_j) \approx 0 \ \forall i, j$ s.t. $i \neq j$, as to not repeat information and dominate over other inputs, and to reduce complexity.

3.2 Data

3.2.1 Tucson Arrests and Neighborhood Datasets

Test:

- **Hypothesis 1**: The type of crime is dependent on: the age of the person committing it; sex of person committing it; the date information (day of the week, month, and time); and relevant neighborhood information (median household income and quantity of houses).
- **Hypothesis 2**: The median household income of an area where a crime was committed is dependent on: the date information (day of the week, month, and time); the age of the person committing it; sex of person committing it; and the quantity of houses in the area.

To get the data to test these hypotheses, use 3 data sets from the City of Tucson:

Neighborhood Income for **2019**, which contains:

- ”NAME”: neighborhood name
- ”MEDHINC_CY”: the median household income for a given neighborhood
- ”WLTHINDXCY”: the wealth index for a given neighborhood
- ”TOTHH_CY”: the total number of houses in a given neighborhood

Tucson Police arrests for both **2020** and **2021**, which contains:

- "NHA_NAME": the name of the neighborhood in which the crime was committed. Maps to the corresponding NAME field from the neighborhood dataset
- "sex": the sex of the person who committed the crime
- "age": age of the person who committed the crime
- "datetime_arr": the date and time the crime was committed
- "fel_misd": the type of crime being committed
  - *M*: Misdemeanor
    - Law requires imprisonment by an entity other than the State Department of Corrections after sentencing
  - *F*: Felony
    - Law requires imprisonment by the State Department of Corrections after sentencing
  - *C*: Civil
    - Non-criminal violation.
  - *P*: Petty Offense
    - Law indicates only a fine is necessary.
  - *S*: Status Offense
    - An act committed by a juvenile that is illegal due to the age alone (e.g. curfew, running away).

3.2.2 Data Preprocessing

Processed the CSV's such that the outputted file contains only the relevant data. This consisted of a few steps:

- Process Arrests Data:
  1. Read the CSV's into dataframes, and drop all unwanted columns.
  2. "datetime_arr" entries are Strings formatted as "yyyy/mm/dd hours:minutes:seconds:UTC_offset". Separate month, day of week, and time fields by partitioning and processing this data, convert it into new "day" (meaning day of the week), "month", and "time" columns, dropping the original "datetime_arr" column.
- Process Neighborhood Wealth Data:
  1. Read CSV into dataframe, drop all unwanted columns.
- Merge the Cleaned Data:
  1. The arrests 2020 & 2021 data is combined into one dataframe.
  2. Then, the neighborhood wealth dataframe is used to map each "NHA_NAME" entry from each row of the arrests dataframe to the neighborhood's associated wealth statistics
  3. Remove "NHA_NAME" column, since it is no longer useful
  4. Remove null/empty entries, and delete any non-numeric/invalid "age" entries
  5. Clean the rest of the cols by deleting empty/useless entries
  6. Write out the cleaned data to a new CSV

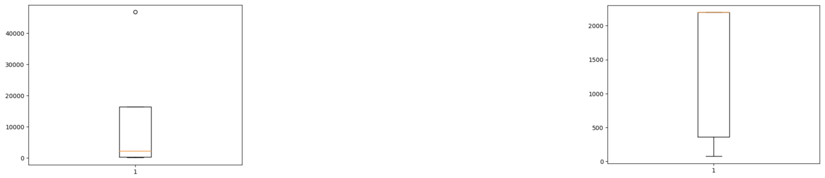
3.2.3 Model Data Processing

The data is now clean. However, it cannot be used to train the models just yet. The ML models chosen are unable to utilize categorical variables as inputs. Therefore, it is needed to convert categorical inputs into numerical ones instead. To do this, one-hot-encode the ”day”, ”month”, and ”sex” fields so that they can now be utilized as inputs.

Additionally, it is needed to normalize some of the inputs so that their scales are comprable with each other. Since, for example, age is usually from about [0,100] and median household income could be from [0, 200000], it is necessary to normalize such fields (wealth index, household income, age, number of houses) so they are all scaled the same.

Finally, it is necessary to resize the data so that certain classes aren't dominated by the others. There are two approaches taken:

1. Randomly remove a subset of the two majority classes to remove the class outliers
2. Statistically generate *new* data points for the two underrepresented class.
   - Find the PDF for the non-normalized features and generate g samples for each
   - Pull $g \times n_n$ samples from a t-distribution (for the normalized features of size n_n
   - Add this data to the dataframe



(a) Original Data Boxplot

(b) New Data Boxplot

Figure 1: Boxplots of fel_misd Classes

Class	Original Size	Final Size
C	46803	2194
M	16407	2194
F	2194	2194
S	240	360
P	50	75

Table 1: Class Sizes

Figure 1a shows there is a clear class outlier in the dataframe. **Figure 1b** shows that the range of the new dataset is much smaller and thus allows for a more balanced distribution of data, resulting in a higher performing model. The original and resulting class sizes are shown in **Table 1**.

The choice for the final class size was as follows: condense the overrepresented class sizes to that of the median. Increase the size of the underrepresented class by 50%. The former was done to force the classifier to learn more than to just guess one class. The latter was done to help decrease the range of class values.

The data is then split from the last dataframe into 80% (training) and 20% (testing) subsets.

3.3 Logistic Regression

All combinations of the input features of length one to five are found and used in five-fold cross validation. The combination with the best resulting accuracy is saved. The training and testing data remove the unneeded features. The logistic regression is then fit with the *multi_class* parameter set to *ovr*.

The trained model is passed the test dataset and the results are stored. The accuracy for both the training and testing data is computed. The results of the test set are saved in a confusion matrix and a classification report. The coefficients for each input feature are also saved.

3.4 Linear Regressions

All combinations of the input features of length one to four are found and used in five-fold cross validation with a Ridge Regression. The combination with the best resulting accuracy is saved. The training and testing data remove the unneeded features, as well as the α_r with the highest R^2. α_r is saved. Twenty-fold cross validation finds the α_l with the lowest mean squared error for the Lasso regression. α_l is saved.

The Ridge Regression is fit with α_r. The trained model is passed the test dataset and the results are stored. The R^2 for both the training and testing data is computed. The residuals are saved. The coefficients for each input feature are also saved.

The Lasso Regression is created with the *alpha* parameter set to α_l. The model is fit with the training data. The trained model is passed the test dataset and the results are stored. The R^2 for both the training and testing data is computed. The residuals and the coefficients for each input feature are also saved.

4 Results

4.1 Correlation

As mentioned in section 3.1.4, the input features should be uncorrelated.

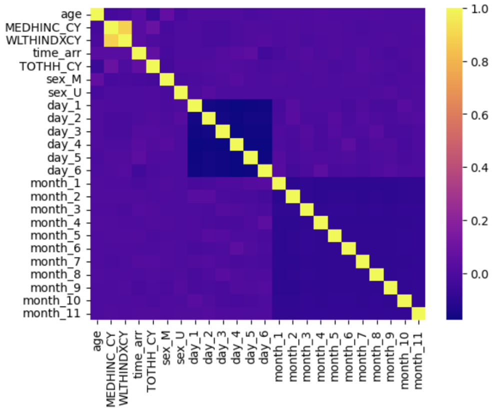
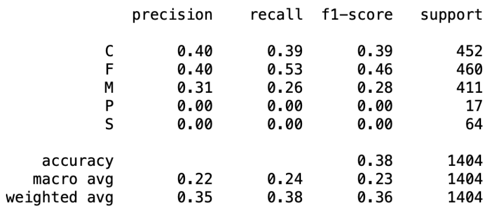


Figure 2: Heatmap of Correlations for Input Features

Figure 2 shows the input features are mainly uncorrelated. The only correlated inputs are the median household income and the wealth index. This makes sense, however, removing wealth index from the inputs will guarantee the inputs features are all fairly uncorrelated.

4.2 Logistic Regression

The results of the five-fold cross validation found the following features to have the best training accuracy: median household income, age, and day of week. The resulting training and testing accuracies were approximately 38.9% and 37.8% respectively.



```
              precision    recall  f1-score   support

           C       0.40      0.39      0.39       452
           F       0.40      0.53      0.46       460
           M       0.31      0.26      0.28       411
           P       0.00      0.00      0.00        17
           S       0.00      0.00      0.00        64

    accuracy                           0.38      1404
   macro avg       0.22      0.24      0.23      1404
weighted avg       0.35      0.38      0.36      1404
```

Figure 3: Classification Report

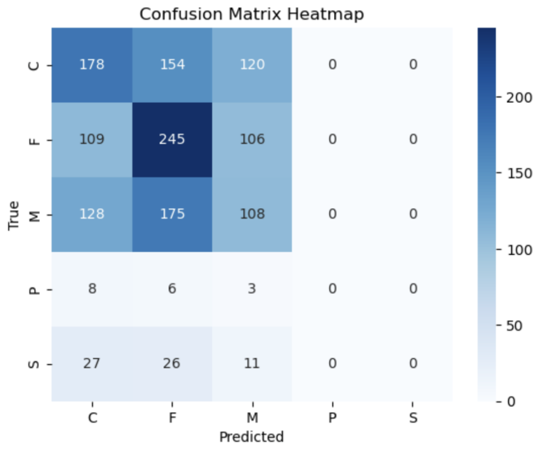


Figure 4: Confusion Matrix

Figure 3 shows all of the metrics were 0.00 for the S and P classes. The weighted averages of each of the metrics is around 0.36, however the macro averages are around 0.23.

Figure 4 shows there are no predictions for the P or S classes. This is likely a result of these being the two classes with small sample sizes. Additionally, the rest of the guesses are concentrated primarily around the C, F, M classes. This is a result of these classes each having the same amount of samples, which was much greater than the other two classes.

Since there are five classes, the probability of randomly guessing one correct is $\frac{1}{5}$. Finding a t-test using $H_0 : \mu \leq 0.2$, $H_a : \mu > 0.2$, and $\alpha = 0.05$ will show that probability of the model being better than chance alone, implying the regression has learned something meaningful from the data.

Using $\hat{p} = 0.378$, $\hat{\sigma}^2 = \frac{\hat{p}(1-\hat{p})}{n-1} = \frac{(0.378)(1-0.378)}{1404-1} \approx 0.000168$:

$$F_{1403}(T) \approx 1 \implies p = 1 - F_{1403}(T) = 0 \tag{9}$$

Since the p-value of $0 < \alpha = 0.05$, this test data is unlikely to happen by chance alone, showing that there is some meaningful information based on the three features to predict the class. In other words, the type of crime is somewhat related to the median income of the area the crime was committed, age, and day of the week. Since the model trained found statistically significant results, the coefficients for the regression can provide insights into the relationship between the features and classes.

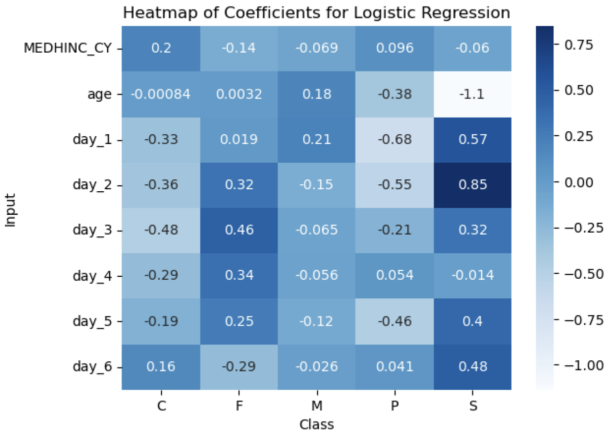


Figure 5: Logistic Regression Coefficient Heatmap

Figure 5 shows class S is highly related to age. This follows logically, since status crimes are only illegal due to an underaged person committing them. Additionally, age has medium strength in both the P and M classes, and is very weak for C and F.

The days of the week have a pretty uniform spread within each class, but vary widely between the different classes.

Lastly, the median income of the area has medium strength in the C and F classes, however is weak in the rest.

4.3 Linear Regressions

4.3.1 Ridge Regression (L2)

The results of five-fold cross validation on the Ridge model found the best performing features to be the sex of the person committing the crime, the day of the week, and the month. Additionally, the cross validation found $\alpha_r = 0$, $\alpha_r \in \{10^{-1}, 10^3\}$. This implies that a normal linear regression is better than a ridge regression.

The train and test coefficients of determination (R^2) were 0.00548 and -0.00463. The test R^2 shows that this model does not accurately predict the median household income for an area.

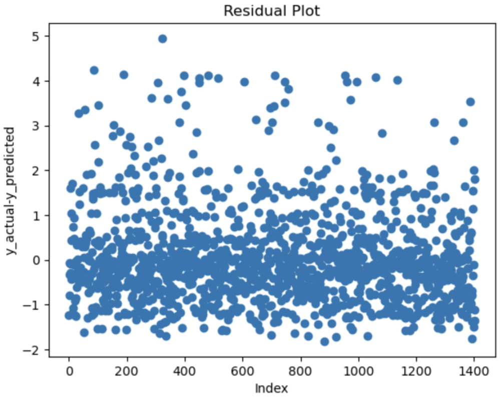


Figure 6: Residuals for Linear Regression

Figure 6 shows that the residuals of the normalized outputs are pretty uniformly spread and are not close to 0. In terms of z-scores, the residuals are extremely distant from 0. This reaffirms the low R^2

4.3.2 Lasso Regression (L1)

The results of twenty-fold cross validation on the Lasso model found $\alpha_l \approx 0.0038$ performs the best.

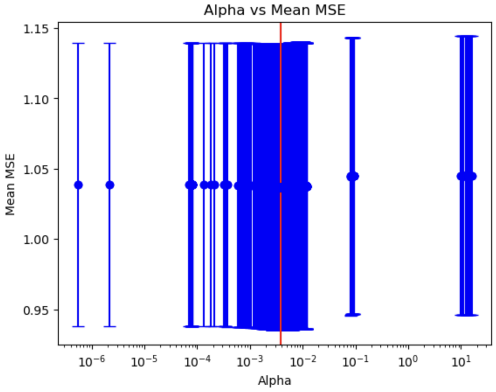


Figure 7: MSE for Each Alpha Value

Figure 7 shows that the best MSE occurs at the red line, at $\alpha \approx 0.0038$.

The train and test R^2 values were 0.00358 and -0.00364. The test R^2 shows that this model does not accurately predict the median household income for an area.

The residuals in **Figure 8** shows that the predicted values are far from zero on a normalized scale. Again, this reaffirms a low R^2.

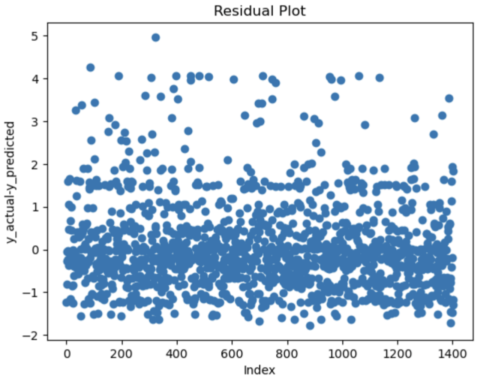


Figure 8: Residuals for Lasso Regression

4.3.3 Coefficients

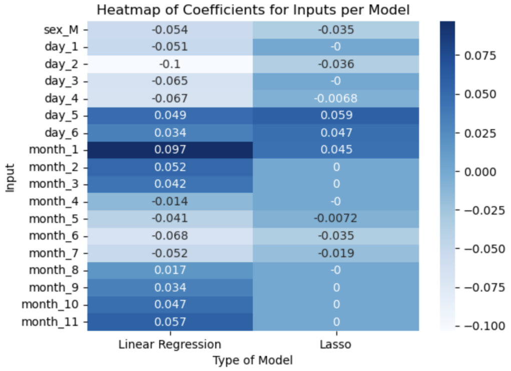


Figure 9: Coefficients for Linear and Lasso Regressions

Figure 9 shows that the Linear Regression has coefficients with an absolute value close to 0: $(|w_i| \leq 0.1 \ \forall i \in K)$. Additionally, this holds for the Lasso as well, with even tighter restriction of $|w_i| < 0.06$. This is a combination of the both the features being uninformative, as well as the L1 regularization coming into affect.

5 Conclusion

The OVR logistic model clearly outlines some relationship between the median household income, age, and day of the week with the type of crime committed in Tucson. With a testing accuracy of 0.378, a p-value of $p \approx 0$ is found. This implies these input features are relevant in analyzing the type of crime. However, the dataset is limited in terms of the P and S classes, so the model fails to perform well for these classes. Despite this, the model was able to pull some useful information in terms of the S model, with the coefficient attached to age. Additionally, the model coefficients show some relationship with the median household income to the type of crime. The largest median household income weight was for the C class of 0.2, implying that the civil crimes may be partially dependent on the wealth of the area. Throughout the entire dataset, however, it seems that the day of the week had the strongest relationship. This could be interpreted as affirming our hypothesis, however an accuracy of 0.378 is concerning to act on.

Both of the linear regressions shows there is not much information regarding the wealth of an area a crime was committed based on the sex of the person committing the crime, day, and month. With very low R^2 values, the models do not represent much of the variance of the median house hold income based on these input features. This contradicts our second hypothesis, as the model failed to correlate the input features to the median house hold income.

This information must be interpreted correctly and carefully. First, the data is limited to 2019 to 2021 Tucson data, meaning the data may not be generalizable beyond this scope. Additionally, using the median household income may pose ethical risks, namely in class bias which could result in increasing current inequalities. In conjunction with this, the data could be skewed by the current policing of Tucson and thus have more crimes in over-policed areas as a result. Also, the results seen may be a result of other factors. An example could be social issues are typically focused and seen more in impoverished areas which could directly lead to more arrests as a result of protesting and other actions. Ultimately, as mentioned before, an accuracy of 0.378 is concerning and may not be extremely impactful in allocating resources in a meaningful way. Thus more research would need to be done to create a better performing model to reaffirm this finding and creating a stronger model. This can be done in multiple ways including: expanding the dataset (less data elimination), better data generation techniques, or more advanced models.

References

Aryan Jadon, Patil, A., & Shruti Jadon. (2024). "A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting. Lecture Notes in Networks and Systems", 117–147. https://doi.org/10.1007/978-981-97-3245-6_9

City of Tucson (2019). "Neighborhood Income" [Data set]. GIS Data. https://gisdata.tucsonaz.gov/datasets/59f033d07eae41b0bdc21db87375d721_0/explore

City of Tucson (2020). "Tucson Police Arrests - 2020 - Open Data" [Data set]. GIS Data. https://gisdata.tucsonaz.gov/datasets/71ee61a5917d4382a423ccfa3d4dfd15_52/explore

City of Tucson (2021). "Tucson Police Arrests - 2021 - Open Data" [Data set]. GIS Data. https://gisdata.tucsonaz.gov/datasets/7c7c881c1fff44ec8a8c2ab612700271_67/explore

Dr.A, Usha Ruby. (2020). "Binary Cross Entropy with Deep Learning Technique for Image Classification." International Journal of Advanced Trends in Computer Science and Engineering 9 (4): 5393–97. https://doi.org/10.30534/ijatcse/2020/175942020

Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions. IEEE Access, 11, 60153–60170. https://doi.org/10.1109/access.2023.3286344

Patterson, E. Britt. 1991. "Poverty, Income Inequality, and Community Crime Rates." Criminology 29 (4): 755–76. https://doi.org/10.1111/j.1745-9125.1991.tb01087.x