DS 4021 Machine Learning Final Report

Corrine Fogarty, Thomas Cusick, Kylie Stephens

Introduction

Our research question was: How does the crime type predict the location where the crime was committed, and are certain communities or areas (location) more likely to experience certain types of crime and or crimes that lead to arrests?

Our motivation behind choosing this dataset and this research question is to better grasp the crime landscape in Chicago. A machine learning analysis on this comprehensive dataset would offer the opportunity for local authorities to allocate resources more efficiently, better advise the community in terms of safety, and help to piece together more effective crime prevention strategies. Data from the real world allows for real world solutions and applications of said solutions. Methods and insights like these can be derived and repurposed to other cities or areas in order to increase safety and minimize crime.

This dataset is from the Chicago Data Portal, and it actively tracks all crime, excluding murder, in the city of Chicago, starting in the year 2001. Because of the immense size of this data set, we queried it to include only crimes from November 15, 2020 to November 15, 2025. This dataset contains the following columns 'ID', 'Case Number', 'Date', 'Block', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', and 'Location'. This data is extracted from the Chicago Police Department's CLEAR system, and the exact addresses are hidden (only shown at block level) to protect individual's privacy.

This dataset did include missing values which we dropped. It was roughly 12,000 out of 940,000 rows, so our group decided due to the sheer size of the data, our models would not suffer from a 1.2% loss. Also, due to how large the data set was, the training times became very long, some upwards of 30 minutes. To resolve this issue, random samples of our data sets were used to train the models on, roughly 100,000 out of 900,000 rows. For every model, since we were trying to find relationships with location, we dropped columns Beat, Ward, Community Area, and District because they would've skewed our results, seeing as they are indicative of and directly related to location. We wanted to focus on the information of the crime itself.

Methods

In our random forest model, our categorical columns were encoded using a label encoder. Columns with high cardinality like 'ID' or 'Case Number' were dropped because they weren't relevant to the relationship we were trying to predict and also used a lot of processing time. For the neural network, the same high cardinality columns were dropped and features were scaled using the standard scaler method. Categorical columns were also encoded with a label encoder (OneHot Encoder). In the SVM, features were scaled using the standard scaler and categorical columns were encoded like other models. The same cat-columns were dropped like other models. In the SVM and random forest models, GridSearchCV was used for cross-validation with 3 folds. In the neural network, kfold cross validation was used with 5 folds. MSE and R^2 were used as the metrics to assess model performance because we were using regression models. For the penalized linear model, 'Description' was also dropped due to feature explosion. Each model was optimized after splitting the training set into an internal training and validation set– lasso, ridge, and elastic net linear regression models were trained and optimized. Ridge produced the lowest MSE and highest R2, although all R2 scores were poor. The Ridge model was optimized with an alpha value of 10 and 5000 iterations. Linear regression had to be used due to the fact that latitude and longitude were continuous values.

Results

Our group found that the best model was our penalized ridge linear model, resulting in a test MSE of 0.005305 and a test R^2 of 0.0410. However, we realize that these results are not telling of a strong relationship between crime and location. After removing geographic identifiers to avoid data leakage (i.e. district and community area would give away location), the linear model achieved an R^2 of 0.0410, indicating that crime characteristics alone (type, domestic status, arrest outcome) are insufficient for precise location prediction. This suggests crime locations are not strongly determined by incident attributes in isolation, and spatial patterns require explicit geographic features for accurate modeling. Using community area or district in place of location may have resulted in a better correlation between crime type and location as it is hard for a model to learn latitudes and longitudes when each continuous value is only occurring once. The final test set metrics for the Ridge Model resulted in an MSE of .005527 and an R2 score of .0386. This is a poor R2 score, and we would not recommend this model to be used for real-world initiatives to allocate resources or advise the community on safety more effectively.

Train Results

	Random Forest	NN	SVM	Lasso	Elastic Net	Ridge
MSE	0.005366	0.005605	0.005442	.005465	.005468	0.005305
R^2	0.0338	0.143099	0.0195	.013356	.012982	.041084

Test Results

```
Final Test Set Metrics for Ridge:
MSE: 0.005527
RMSE: 0.074344
MAE: 0.059620
Latitude RMSE: 0.086284
Longitude RMSE: 0.060076
R2 Score: 0.038555114332885776
```

Discussion

Our results could suggest that crime monitoring tools that target populations, demographics, and regions may not be reliable because for at least this dataset, no reliable relationship was found. The dataset lacked features that would be truly indicative of crime and location, such as socioeconomic factors, temporal trends, land-use, zoning maps, and or infrastructure– these would most likely provide a stronger correlation. As far as the random

forest and SVM, the random forest likely performed better because of the size of the data set. Random forests also handle noisy or highly dimensional data better and work well with mixed data types, and our dataset has a lot of strings. NN would have required more signal and cleaner data than this dataset provided. The ridge model may have performed the best because it is fast and scales to large datasets well; it handles noisy, high-dimensional features better than the other models tested. A penalized ridge regression also has a lower variance and is less likely to overfit data.

We experienced a few limitations for this project. First of all, we had to remove many columns for every model because they were too closely related to location and would skew our results. Therefore, we had less features for our model than we originally thought we would. Some of our columns also had a lot of unique values and therefore had to be removed from the dataset because the processing time was too long and encoding the variables led to more complications with the data. Realistically, after running our models, our group realized it would have been more effective to use community area as our target variable. This is because location was based on coordinates and we had to split and transform the location because the coordinates were stored in one string. For some models, we also had to use random sampling to get a picture of the dataset because it had so many rows and would take too long to process; it also caused feature explosion due to categorical features. The community area was an integer in the dataset and we may have been able to find a better relationship if we used that. The biggest limitation in this project lies in what we chose to predict– location. While our goal was to predict location (and to determine if crime type could predict location of where the crime occurred), we should have chosen community area or district as location was (latitude, longitude) – these are continuous values and are extremely hard to predict, making it almost impossible for the model to provide meaningful results (for the real world). Additionally, using a linear model is not effective for spatial predictions; however, because logistic regression requires a discrete, categorical dependent variable, it cannot produce real-valued coordinate predictions. Applying logistic regression here would incorrectly force geographic positions into artificial categories, losing the precision and structure of the original problem. Overall, the models and strategies used for EDA, data cleaning, optimization were strong, but the dataset and specific features and outcome we chose to predict were weakly correlated and did not provide meaningful results to answer our original research question.