Machine Learning Engineer Nanodegree

Capstone Proposal

Wesley Strange
January 24th, 2019

Proposal

Domain Background

For my final project I will be using Chicago Crime data from years past to create a Machine Learning model to predict when violent crimes are most likely to happen. I believe the Police Department could benefit from this model by being able to better allocate their resources in attempt to prevent some of the more violent crimes that are being committed.

I live in the Northwest Suburbs of Chicago, so I would personally like to see the crime rates in Chicago reduced. I would like to feel safe when I'm out exploring the many beautiful sights that Chicago has to offer. I also chose this domain since Chicago has received a lot of negative media coverage in recent years due to the high number of murders and gun violence. According to the FBI crime dataset, in recent years Chicago has experience violent crime rates nearly three times higher than the national average (Uniform Crime Reporting (UCR) Program, 2018).

This type of problem has been solved before using Machine Learning. In the article "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data" they attempted to predict crime using a variety of data sources, including demographic and mobile phone data (Andrey Bogomolov, 2014).

Problem Statement

The problem is that there are too many violent crimes being committed in the Chicago neighborhoods. According to the FBI crime dataset, in recent years Chicago has experienced violent crime rates nearly three times the national average (Uniform Crime Reporting (UCR) Program, 2018). In 2016, the national average was 387 violent crimes per 100,000 residents while Chicago clocked in at 1,105. In 2017, the national average was 383 while Chicago was at 1,099 per 100,000 residents.

This is a supervised classification problem. The input features will include a lot of data gathered about Chicago (crime, traffic, weather, professional sporting events/results) and some general data (bitcoin price, stock market price, and holidays). The output will be one of two predictions: 1 – the day is predicted to be a high-crime day or 0 – the day is not predicted to be a high-crime day.

Datasets and Inputs

The biggest question I faced was what data to feed into the Machine Learning model. I had a hard time determining what factors might come in to play to cause someone to commit a violent crime. I settled on the following data: Chicago crime, traffic, weather, professional sporting events/results, bitcoin price, stock market price, and holidays.

Chicago Crime Dataset (2001 - Present) **- Target**
This dataset contains the historical crime data for Chicago. I will only be using the data from May 2015 – Nov 2018. I will filter out all the non-violent crimes from the dataset, since I'm only interested in the violent crimes. I will aggregate the data grouped by day to get the sum of the number of violent crimes happening each day. I will create a new categorical variable called "High_Volume_Day" that is used to

identify if a day qualifies as a high-volume violent crime day. This new variable will be the target value that we will be trying to predict. I will set a threshold so that approximately 25% of the days are identified as high-volume days. If the field value is 1 then it will be considered a high-volume violent crime day. Conversely, if the value is 0 then it will not be considered high-volume day. The distribution of the classes will not be balances, since I'm purposely limiting the number of high-volume days to about 25%. There will be about 1,200 records once the data has been aggregated by day.

The data was obtained from the Chicago Data Portal website <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g>.

Chicago Weather (May 2015 - Present) **- Features**
I believe weather could come into play in a couple of different ways. First being peoples' overall mood due to the weather. Personally, I'm usually in a more upbeat mood when the weather is warm and sunny. On the flipside, when the weather is cold, rainy, or snowy, I'm more likely to be easily irritated. The second way I could see this coming into play is that people are more likely to be outside if the weather is warm and sunny, so they are probably more likely to interact with other people which might give them more opportunities to commit a violent crime.

Numerical features include: Air Temperature, Wet Bulb Temperature, Humidity, Rain Intensity, Interval Rain, Total Rain, Precipitation Type, Wind Direction, Wind Speed, Maximum Wind Speed, Barometric Pressure, Solar Radiation. There are several records for each day from multiple locations, so I will have to summarize the data into one record for each day.

The data was obtained from the Chicago Data Portal website <https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75>.

Chicago Sports Data **- Features**
It's hard to find anything that interests Chicagoans more than sports. Often you can tell the outcome of last night's big game by the morale in the office the next morning. Rarely a day goes by when I don't get stopped to chat about the Cubs chances in the upcoming season. Hence, I believe sports could play a role in the occurrence violent crimes.

Categorical features: Opponent, Home/Away, Result

The data was obtained from the Sport Reference suite of websites <https://www.sports-reference.com>.

Dow Jones Historical Data **- Features**
Money makes the world go around. When the market is booming, people are probably more likely to be in a good mood. When the market is slipping, people are more likely to be on edge. I can see this playing a big role.

Numerical feature: Gain/Loss (new value that will be calculated)

The data was obtained from Yahoo Finance <https://finance.yahoo.com/quote/%5Edji/history?ltr=1>.

Bitcoin Historical Data **- Features**
Same explanation as the Dow Jones explanation.

Numerical feature: Gain/Loss (new value that will be calculated)

The data was obtained from Coin Market Cap website <https://coinmarketcap.com/currencies/bitcoin/historical-data>.

National Holidays **- Features**
I believe people are more likely to be off work on days that are considered a National Holiday.

Therefore, they're more likely to engage in social gatherings (parties, etc.) and drinking which could lead to more violent behavior.

Categorical feature: Holiday?

The list of recognized National Holidays was obtained from Wikipedia <https://en.wikipedia.org/wiki/Federal_holidays_in_the_United_States>. The actual dates were obtained by searching in Outlook Calendar.

Solution Statement

The solution I'm proposing is to use a Machine Learning model to identify when violent crimes are most likely to occur. The ML model will take as input Chicago traffic, Chicago weather, Chicago sports, Financial Market data, whether it is a Holiday, and it will determine if violent crimes are more likely to happen that day. Having this information available would give the Police an opportunity to allocate more of their resources during those times when the violent crimes are more likely to occur in effort to prevent violent crimes from being committed or if crime has been committed, they will have a better chance of apprehending the criminal.

The supervised learning models that will be considered are Gaussian Naïve Bayes, Decision Trees, Ensemble Methods, K-Nearest Neighbors, Stochastic Gradient Descent Classifier, Support Vector Machines, Logistic Regression. Since the dataset is imbalanced, 3:1 in favor of not a high-volume crime day, I will be using the AUROC as the metric to evaluate the effectiveness of the ML model (Lador, 2017).

Benchmark Model

I'll evaluate the model based on whether it gets a result better than random choice. I'll refer to this benchmark model as the "Naive Predictor". The Naïve Predictor will be the value we get if we chose a model that always predicted '1' (i.e. the day is considered a high volume violent crime day).

For the ML model to be useful, we'll need to limit the number of days that are considered "High Volume" otherwise the Police would be on high alert too often and the model wouldn't be useful. For this project, I'll limit the threshold of high volume days to no more than 25%. To identify whether a day is a high-volume day I will sum up the number of violent crimes by day and then create a threshold that identifies approximately 25% of the projects as high-volume days.

Evaluation Metrics

AUROC (Area Under the Receiver Operating Characteristics) will be the main metric used to evaluate the effectiveness of the ML model we create for this project, since the dataset is imbalanced as explained in the previous section. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

Measure definitions (Narkhede, 2018):

TPR (True Positive Rate)

$$\text{TPR} = \text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$$

FPR (False Positive Rate)

$$\text{FPR} = \text{False Positives} / (\text{True Negatives} + \text{False Positives})$$

Project Design

Explore the data

I will explore the data to get an idea of what the data looks like. I will use histograms to Identify any skewed distributions that need to be transformed prior to inputting the data into the ML model. I will also explore the data to identify if there are any potential gaps or issues with the data that need to be addressed before proceeding further.

Prepare the data
I will filter out all the non-violent crimes from the dataset, since I'm only interested in the violent crimes. Then, I will aggregate the data grouped by day to get the sum of the number of violent crimes happening each day. I will create a new feature called "High_Volume_Day" that is used to identify if a day qualifies as a high-volume violent crime day. I will set a threshold so that approximately 25% of the days are identified as high-volume days. If the field value is 1 then it will be considered a high-volume violent crime day. Conversely, if the value is 0 then it will not be considered high-volume day. Next, I will transform any of the skewed continuous features that were identified during the data exploration process. Then, I will normalize the numerical features so that they are all equal to values between 0 to 1. Then, I will use one-hot encoding to get all the categorical features into a format that can be input into the ML model. Finally, I will split the data into training and testing sets.

Establish Benchmark Model Performance
The Naïve Predictor will be the value we get if we chose a model that always predicted '1' (i.e. the day is considered a high-volume violent crime day). The AUROC metric will be captured and used to evaluate our ML model's performance later.

Apply Supervised Machine Learning Models and Evaluate Model Performance
The supervised learning models that will be considered are Gaussian Naïve Bayes, Decision Trees, Ensemble Methods, K-Nearest Neighbors, Stochastic Gradient Descent Classifier, Support Vector Machines, Logistic Regression. I will pick a few of the models as appropriate and apply each model to the dataset. I will review the results of each model to determine which model's performance was the best. Then, I will proceed to see which model performed the best. I will then use that model to improve upon.

Improve Model
I will use GridSearchCV to fine tune the chosen model in hopes to achieve even greater performance.

Final Model Evaluation
I will compare the AUROC results of the optimized model to that of the Naïve Predictor.

Extract Feature Importance
Lastly, I will extract the feature importance values to determine which features are providing the most predictive power.

Bibliography

Andrey Bogomolov, B. L. (2014, 09 10). *Once Upon a Crime: Towards Crime Prediction.* Retrieved from arXiv: https://arxiv.org/pdf/1409.2983.pdf

Lador, S. M. (2017, 09 05). *What metrics should be used for evaluating a model on an imbalanced data set?* Retrieved from Towards Data Science: https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

Narkhede, S. (2018, 06 26). *Understanding AUC - ROC Curve*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Uniform Crime Reporting (UCR) Program. (2018, 01 26). Retrieved from FBI: https://www.fbi.gov/services/cjis/ucr