Minority Report: Predicting Crime Rates for Neighborhoods Using Local Infrastructure and Socio-economic Indicators

Ajinkya Indulkar
CICS
University of Massachusetts Amherst
Amherst, MA 01003
aindulkar@cs.umass.edu

Anish Pimpley
CICS
University of Massachusetts Amherst
Amherst, MA 01003
apimpley@cs.umass.edu

Kartik Chhapia
CICS
University of Massachusetts Amherst
Amherst, MA 01003
kchhapia@cs.umass.edu

Abstract

In this project, we propose a novel approach to predict crime for neighborhoods in a city from multiple sources, in particular nearby landmarks, census and demographic data. Prior work on crime prediction has primarily relied on historical crime data and has revolved around crime hot-spot detection. We present a preliminary investigation of infrastructure-based crime prediction. We tested our model on 3 categories of crimes : violation, misdemeanor and felony. Our experimental results significantly outperform a baseline linear model. We hope that this project will help decision makers and city officials allocate resources in areas that have undergone significant changes in infrastructure and demographics. Furthermore, we also highlight correlation between various landmarks, socio-economic indicators and crime.

1 Introduction

Crime is a vice that has plagued our society since the beginning and is bound to stay forever. Criminal psychologists and sociologists have been studying crime for a long time[1][2][3]. Reducing crime is not straightforward and making reforms to reduce crime may often result in increased crime[4]. Studies have shown relation between economic indicators and crime[4]. Studies have indicated that there is a direct causal relationship between education and crime[5]. Considering these two results, our hypothesis is that local spatial infrastructure and local demographics play an important role in criminal behavior. We use these data to predict crime in a particular area or neighborhood.

The motivation behind this study is to help law officials and decision makers in allocation of limited law-and-order resources over a geographic area. For example, in case of a scarce police force, it would make more sense to deploy police patrols and officers to areas that are highly susceptible to crime. This could also decrease response time. Numerous states in the US including California, Washington, South Carolina, Arizona, Tennessee, and Illinois have adapted predictive policing systems[6].

In this project, we would like to predict crime rates for neighborhoods using local infrastructure and socio-economic indicators. This can be modeled as a regression problem in which we try to predict a value that will be a measure of the crime rate of a given neighborhood. There could be several applications of such a model. The regressor could serve as a tool for law enforcement to preemptively gauge expected criminal activity in an area with recent changes in demographics and infrastructure and accordingly allocate resources to particular neighborhoods. Another application of such a model could be to predict the crime rate in smaller cities and towns that do not have much crime rate data recorded but have information about the landmarks such as public places, restaurants, alcohol shops and census data such as number of people living in a neighborhood, median income etc. Furthermore, we also try coming up with a list of factors that have the most influence on crime rate in an area.

Our main contributions with this project will be:

1. Using local infrastructure data and demographic data in predicting crime in an area using state-of-the art machine learning models.
2. Analysis of the performance of a trained model to predict crime in unobserved areas.
3. Discussion of landmark and socio-economic feature importances and their correlation with crime.

2 Related work

In the past few decades, researchers have spent significant time and effort in studying crime behavior. In particular, much work has been done in predicting crime hotspots based on historical crime data. Because of this, decision makers and law enforcement have gained significant insight into crime behavior given time series information of past crimes. These statistical methods intend to help law officials identify geographical patterns in criminal behavior and underlying causes of such crimes. This field has gained immense traction in the past few years has been assigned a term: *Predictive Policing*.

Bogomolov *et al.*[7] have tried predicting crime in a geographic space using mobile phone and demographic data. They found that human behavioral data obtained from mobile network infrastructure along with spatio-temporal demographic data can be used to predict crime with an accuracy of 70%. They framed it as a classification problem and classified points in a geographic area into crime hotspots and non crime hotspots. It might be interesting to see how mobile data would have affected our results. But we leave that for future work as it is beyond the scope of this project. In our project, however, we try to go further and instead of framing it as a classification problem, we frame crime-prediction as a regression problem which will not only indicate which areas have high crime, it will also give us a measure of the expected crime. Chainey *et al.*[8] studied the utility of hot-spot mapping techniques for crime prediction. They claim that there has been very little research in quantifying the accuracy of crime-prediction techniques. They also noted that certain crime types have a positive correlation to crime hot-spots. In this project we do not look at different crime types and leave it for future work. We do, however, try to find correlation between different spatial indicators and criminal incidents.

Wang *et al.*[9] proposed a crime pattern detection algorithm named *SeriesFinder* where they tried to find similarities between the modus operandi of offenders. Using these similarities they were able to find patterns in the modus operandi of crimes using clustering algorithms such as *Hierarchical agglomerative clustering* (HAC) and *Iterative nearest neighbor classification*. SeriesFinder was able to find patterns across crimes that crime analysts had missed. They were able gain deeper insights into crime behavior and modus operandi of offenders. Wang *et al.*[9] concluded that although SeriesFinder was able to correctly cluster related crimes several times, it is not the final answer to all crime and much work remains to be done. Eck *et al.*[10] propose using K-Means clustering and Hierarchial clustering to create spatial ellipses to group areas with similar crime. They admit, however, that crime hot-spots are not necessarily spatial ellipses and therefore a higher-capacity model would would be a better fit for detecting hot-spots. They also use Kernel Density Estimation to learn a two-dimensional probability distribution over the geographical area. They also try to map crimes to census tracts and outline the limitations of using census tracts as a unit. The varying size and shape of census tracts pose a problem in prediction and can often mislead the user in where a hot-spot may exist. This result affects our choice of granularity. Using census tracts would have been easier as the demographic data we use is available for census tracts. But because of the aforementioned limitations, we partition the geographic area into squares as done by Bogomolov *et al.*[7] and Ect et al[10].

2.1 Social-media based crime prediction

Recently, because of the rise in social media, there has been an increase in research in the area of crime prediction using social media. The hypothesis is that the spatio-temporal data attached to a tweet provide some information about crimes. Data from social media also allows one to use context surrounding the social media landscape. Twitter claims that millions of tweets are posted daily by people.

Wang *et al.*[11] investigated methods of using Twitter for crime incident prediction. They used semantic analysis and natural language processing to gain information from Tweets. Wang *et al.*[11] used latent Dirichlet allocation (LDA) for dimensionality reduction and linear modeling for prediction. Traditionally, people had used historical crime data and other simple demographic data to learn patterns in crime. Wang *et al.* found that their results outperform a baseline model that predicts hit-and-run incidents. But the did not address several aspects surrounding crime prediction based on social-media. The investigation was limited to tweets from a news angency and it was only limited to hit-and-run crimes which form a small minority of the total crimes committed in a city. Although they identified that Another potential limitation of this study was that the high volume of tweets may prove to be a bottleneck in training time and it may not be possible to use them with a model that doesn't support online learning. Furthermore, the tweets used by them did not have spatial data which could identify a user's location. Therefore, they were unable to find any correlation between crime in geographic areas and twitter messages.

Gerber[12] in this previous work proposed augmenting existing crime prediction methods with Twitter data. Gerber showed that for 19 out of 25 crime types, the addition of Twitter data improved the performance of crime prediction. Although, there as a major bottleneck in the topic modeling phase which limited the number of tweets that could be used. Given the sheer number of tweets posted on a daily basis, it would not be surprising to gain criminal behavior insight from Twitter data.

A major limitation in all these approaches is that they use historical data to predict crime in a particular area. The problem with this is that crime can only be predicted in areas where we already have significant historical data i.e. they lack portability. Often there are new areas where crime has been traditionally non-existent but due to recent changes in the infrastructure landscape and other demographic factors it is bound to change. Furthermore, there has been no research on finding the correlation between infrastructure and crime. For instance, it is very unlikely that there are many crimes near police stations. A lot of the prior research also doesn't address the role of socio-economic indicators and demographic information in crime behavior. In this project, we aim to address the crime-detection problem from a different direction. We use historical crime data and augment that with a rich feature-set consisting of landmark data and demographic data from the Decennial Census and the American Community Survey. We train machine learning algorithms on this data to predict crime for geographic data. Our model will not be limited to an area where we have observed crime. Once trained, we can use our model to predict crime in any area just based on infrastructure data and demographic data thus removing the need for historical data for that area.

3 Data sets

Since we propose to do crime prediction using several features such as the socio economic factors of people living in the neighborhood and landmarks such as public places, liquor shops, restaurants, schools to name a few we had to devise our own code and methodology to collect the data. As mentioned earlier, we are focusing our efforts on New York City. We have picked up several data points that lie within one of the five boroughs of the city of New York namely Bronx, Brooklyn, Manhattan, Queens, Staten Island. The entire process of data collection is described below. Our first step was to fetch the latitude and longitude of several localities that lie within the city of New York. We created an imaginary rectangular grid that enclosed the entire city of New York including the five boroughs. We had approximately 10,000 unique coordinates of locations that we could use as data points. However, due the geographical shape of the city, about a little greater than 50% of these points that lied inside the bounding box did not lie within the city and thus, these points could not qualify as data points. Specifically, we had 4846 points that were present within the city of New York. To check this, we used a Census Block conversion service provided by The Federal Communications Commission (FCC) [1] The latitude and longitude of these points were noted. Each of these points were situated at a distance of approximately one mile from the closest immediate neighbor. Once we had obtained the latitude and longitude of several points within the city of New York, we focused on obtaining landmarks such as police stations, universities, schools, public places, liquor shops, restaurants that lied in the vicinity of these points. We define vicinity to be a radial distance of 0.8 miles. Since each location point lies at a distance of one mile from each other, our method of selecting landmarks results in neighboring points having some overlap in the landmarks. We do this intentionally because we believe that landmarks that lie in the vicinity of a location should also be included as landmarks for the neighboring points. We have grouped the landmarks under 12 broad and distinct categories. The categories are as follows: alcohol, places of worship, shopping, food, public transport, medicine, public places, police station, university, bank, nightclubs and school. Each of these categories had several sub-categories of landmarks ; for example the alcohol category comprised of bars and liquor shops, the public places category consisted of public places such as airports, amusement parks, aquarium, art gallery, museums and other such landmarks. For each category, we used the number of such landmarks that lied inside the radial distance of 0.8 miles as a feature. Thus, if for a particular location, there were 10 bars and 8 liquor shops that were present within a vicinity of 0.8 miles, then we could use 18 as our feature for the alcohol category. We obtained this information from the Google Maps Places Web API.

For demographic data, we used two sources: Decennial Census data 2010[13] and American Community Survey (ACS) 1-Year Data (2011-2015)[14]. These data are provided for free by the U.S. Census Bureau. But, because of the impenetrable nature of the official census API and the opaque documentation, we ended up using unofficial API provided by Census IRE[2] for the Decennial 2010 Census data and Census Reporter[3]. These unofficial APIs construct thin wrappers over the official APIs, thus offering much more consistent APIs and better structured results. These APIs, however, had very high response times and data collection would have taken several hours. So we used asynchronous parallel requests to collect the Decennial Census data and the ACS data. The smallest granularity available for demographic data is at the level of Census Tracts. But we divided the geographic area into regular equal-sized squares instead of Census Tracts because the varying size and shape of census tracts pose a problem in prediction and can often mislead the user in where a hot-spot may exist as outlined by Eck *et al.*[10]. After dividing the city into evenly sized blocks, we used the centroid of the block to check if it lies within a census tract and assigned data from that tract to the block. This resulted in multiple blocks having the same Census data because of the low granularity but this still gives us much valuable information regarding crimes.

Next, we obtained the information about the number of crimes in the neighborhood. For this purpose, we made use of the freely available dataset by the New York Police Department. This data set consists of 5.2 million crimes that were reported and recorded to the New York City Police Department. Based on numerous law categories and subsections, crimes in this dataset were broken up into 3 major categories: Felony, Misdemeanor, and Violation crimes. The dataset consisted of crimes from the year 2005 to the year 2015. For the purpose of our project, we considered crimes

[1] https://www.fcc.gov/general/census-block-conversions-api-v101

[2] http://census.ire.org/

[3] https://github.com/censusreporter/census-api

between the years 2013 to 2015. Between these years about a million crimes (sum of felony, misdemeanor, and violation crimes) were reported and each crime had a location field that consisted of the latitude and longitude of the location where the crime had occurred. Our methodology for utilizing this crime data was as follows. For each of the 4846 data points within the city of New York, we use the crime data set to find all locations that lie within a distance of 0.8 miles from the original location point. Once we had found these lists of points, we simply note the number of crimes of each category that have occurred and sum the values up. A point to note is that we compute the sum for the crimes in each category and compute the total number of felony, misdemeanor and violation crimes that have occurred. Since our trimmed data set consists of crimes that have occurred over a period of 2 years, what we have essentially computed is the number of crimes that have occurred for each category over a period of two years. Once we have all the above information, for a particular location specified by the latitude and longitude, we combine the landmarks data for that location with the census data and the number of crimes that have occurred. We call this our feature vector for a particular location. The number of features that we have obtained is 86. The entire code for the data extraction is written in Python. The Google Maps Places Web API is used for the process of extracting landmarks and the census API for the purpose of getting the census information.

4 Methodology

4.1 Crime rate as a regression task

We cast the crime rate prediction problem as a regression task. For each evenly spaced coordinate inside NYC, we predict the number of crimes over a 2 year period.
The problem has commonly been modeled as a multi-class classification problem, by binning the crime values and stratifying them into risk levels. It allows for usage of easily interpretable performance metrics such as mis-classification error and direct comparison with non-ML based methods. However, since we bin the regression values into classes, we are effectively relaxing the regression problem into an error tolerant problem model with each label ranging over a large range of crime values. Thus, losing some richness of crime data in the process.
Extracting valuable insight into correlations between features and the crime rate, is a major facet of our project. To avoid loss of feature importance information in converting natively continuous target values into discrete labels, we preferred to model predicting crime rate as a regression task instead.

4.2 Data preprocessing

Utilizing the aforementioned data collection process, we obtained 4845 evenly spaced data points inside the silhouette of New York City. The data made available by the sources we queried from, had not been conditioned with machine learning applications in mind. Thus, preliminary data cleaning was necessary.
Few areas inside New York City had either no reported population or houses. Data points in such regions were deleted from the dataset, as they would have served as noisy outliers. 885 such points were identified and removed. Features with sparse data and 'Nan' results were also removed from the dataset.
Census statistics with regards to age, sex, race and home ownership were reported in raw values, instead of the more preferred percentage metric. To account for traditionally used percentage based census values, 26 new features were engineered. We expect the percentage based values to be independent of existing features such as total population, and allow better independence between features. Each feature is scaled with respect to the feature type's corresponding total. There is some redundancy in the newly engineered features, however we expect it to be eliminated in the feature selection step.

The new features are as follows:

- **Housing**: % houses vacant, %houses occupied,
- **Employment**: %unemployed, %in labor force, %employed,

- **Family**: Families per capita, Families per housing unit, Houses per capita, %owned clear, %owned w. mortgage, %renter occupied, %non family households.
- **Race**: %black, %white, %asian, %other,
- **Sex**:%F, %M, Sex ratio
- **Sex by Age**: F/M under 10, 10-17, 18-24, 25-35, 35-60, 60+

Most of the models we experiment on, are sensitive to scaling. Thus, we scale all of our features with respect to mean and unit variance. We keep our target values untouched for interpretability. The final dataset contains 3960 data points and 110 features. A train-test split of 0.7-0.3 was chosen. Random seeds for all possible values in our project were set to 0, for better reproducibility.

4.3 Feature Selection

To eliminate redundancy in the dataset, we implemented a model based feature selection method in Lasso to identify linearly correlated features. Lasso is a linear model trained with an L1 prior as the regularizer. The objective function for Lasso is as follows:

$$(1/(2 * n_{samples})) * ||y - Xw||_2^2 + alpha * ||w||_1$$

The objective function does not have a closed form solution and iteratively converges to a global optimum. The $L1$ constraint region in a bivariate case can be visualized as follows:

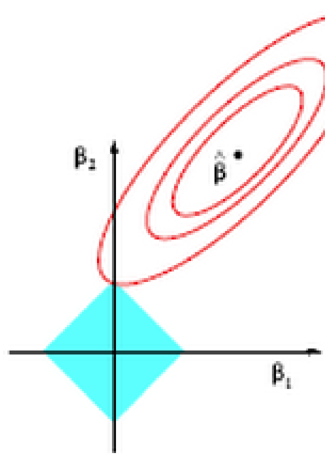

Figure 1: Lasso L1 constraint region

As can be seen from optimum at the intersection between the regulizer and OLS objective function, Lasso drives coefficients of linearly correlated features to zero. We utlize this property to identify redundancy in our dataset, and eliminate features with zero coefficients.
The feature selection process selected a reduced subset of 67/110 features.

4.4 Model Selection

We prioritized high interpretability, adequate capacity and availability of coefficients, during the selection of a regression model. We trained various models, each with unique strengths and varied objective functions.
Ordinary Least Squares (OLS) was used as baseline regressor, with high interpretability and minimal tuning requirements. The closed form solution offers high reproducibility and is quick to both train and test.
Ensemble methods have been shown to achieve class leading performance and have been the go to regressor in past crime based studies. Random Forests (RF) and Gradient Boosted Trees (GBT)

provide significant reduction in variance, while preserving the interpretability and readily providing feature importances. Both RFs and GBTs use Regression Trees (RTs) as their base regressor. In both cases, an ensemble of piece wise rectilinear RTs as weak learners, are used to effectively model a non-linear strong learner.
The 'number of estimators', 'max depth' and 'number of features in each RT', are the primary hyper parameters for both RFs and GBTs.
In RFs, each tree in the ensemble is built from a sample drawn with replacement from the training set. In addition, when splitting a node during the construction of the tree, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases but, due to averaging. Its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.
Gradient Boosted Regression Trees (GBT) is a generalization of boosting to arbitrary differentiable loss functions.
GBT builds the additive model in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

At each stage the regression tree $h_m(x)$ is chosen to minimize the loss function given the current model F_{m-1} and its fit $F_{m-1}(x_i)$
Gradient Boosting solves the minimization problem numerically via steepest descent. The steepest descent direction is the negative gradient of the loss function evaluated at the current model F_{m-1} which is evaluated for the least squares loss function.

$$F_m(x) = F_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i))$$

In cases of higly non-linear mappings decision tree based ensemble methods might not be able to fully represent the crime rate. In such cases high capacity models would be necessary. We chose 2 contrasting regressors in Support Vector Machines (SVM) and Neural Networks (ANN) to test against the ensemble methods. SVMs serve as great off-the-shelf regressors, that work well with medium sized datasets. RBF kernel SVMs are easy to tune with only 2 primary hyperparamters in C and γ. The epsilon insensitive loss function for SVM mimics the margin property observed in SVM classification and is unique to support vector machines. Neural Nets on the contrary, often require large datasets and are much more difficult to tune. However, they can learn their own nonlinearity and have shown to achieve branch leading results in the past few years. Both suffer from the issue of not having readily availible feature importances and being analogous with black box methods in terms of interpretability. The SVM was tested for an RBF kernel. The Neural net models were tested for both 1,2 hidden layers. Both Neural net models were simple multilayer perceptron models. Further details on the SVM, ANN hyperparamter ranges can be found in the experiments section. Because of the aforementioned shortcomings of SVM and ANN, we wish to avoid them unless they outscore both the ensemble methods.
The MAE scores for both ANNs and SVM were either identical to or higher than those achieved by both RFs and GBTs. Thus, we can safely assume that RFs and GBTs are able to capture patterns in our problem as competently as any high capacity model. RFs and GBTs also significantly outscore the baseline OLS method. Thus, we select the 2 methods for building our final model.

4.5 Hyper parameter selection and cross validation

In all of the above 5 models, we utilized 3 fold cross validation. As compared to random sampling methods, every data point gets used at least once. For a larger number of folds, we risk having sparse validation sets that are individually biased and incomplete representations of the original dataset. At K=3, an ideal trade off between density and bias was achieved. We conducted grid search over linear ranges to identify optimal hyper parameters for most models. Thus, for regressors such as RFs and GBTs with smaller parameter spaces, we traded off exploration for exploitation. For SVMs we improve exploration by conducting grid search over an exponential range of γ and C. For the significantly larger hyper parameter range of ANNs, we use Randomized search for maximum exploration.

4.6 Platform and libraries

The project was build with Python 2.7. Models were imported from the SKlearn library. Pandas, Numpy and Scipy were used elsewhere for data handling and mathematical processing. All experiments were run on standard personal laptop setups.

5 Experiments and Results

We have chosen Mean Absolute Error (MAE) as the performance metrics for the regressors that we experiment with. RMSE values for each regressor were also compute, however RMSE was not used as the loss function. For visualizations we will be reporting the MAE. As compared to MAE, RMSE being a squared error metric adds emphasis to outliers, while placing much lower importance smaller error values. We much prefer the even handed approach of MAE over the out lier emphasizing RMSE loss function. The MAE values for the 5 regressors were compared for the following subsections of the data:

- Census only
- Landmark only
- Both

Segmenting the features into landmark and census data, goes back to the primary objective of our study. It allows us to gauge the effectiveness of one source of data on its own and gain invaluable information, when both feature sets are used in tandem. Following are the MAE values for each regressor for every aforementioned Feature set:

Table 1: MAE of different regression models for different subsets of data

	OLS	GBT	RF
Landmarks	390	294	304
Census	608	378	378
Both	372	242	243

In Table 1 we highlight the results that are obtained by running Linear Regression (OLD), Gradient Boosted Trees, and Random Forest models on our dataset. In our first experiment represented by the first row in the table, we show the results obtained by running the three different regressors on the dataset with only the landmarks information taken as features and ignoring the census data. In our second experiment, we take only the census data as features and ignore the landmarks data. In our final experiment, we include the complete dataset and consider both the landmarks and census data as features for our model.

From the table, we observe that the mean absolute error obtained by using only the landmarks features outperforms the mean absolute error that is obtained by using only the census features. This leads us to infer that the landmarks features are richer than the census features and contribute significantly in predicting the number of crimes that will occur in a neighborhood.

Table 2: Comparison of different models on the full dataset

	MAE: (Landmark + Census) data
OLS	372
GBT	242
RF	243
ANN	315
SVM	232

In Table 2 we report the results obtained by running the entire data set that includes the census and landmarks features with various machine learning models. The models that we have tried are the following: Linear Regression (OLS), Gradient Boosted Trees (GBT), Random Forest (RF), Neural

Networks (ANN), and Support Vector Machines (SVM). We can observe that SVM is the best performing model since it has the lowest mean absolute error value. GBT and RF have almost similar performance as the SVM model. For further analysis, we choose the GBT and RF models because the sci-kit learn library that we are using returns the feature importance of each feature for these 2 models. Along with those models, we also analyze the OLS model because this serves as a baseline for us to compare the performance.

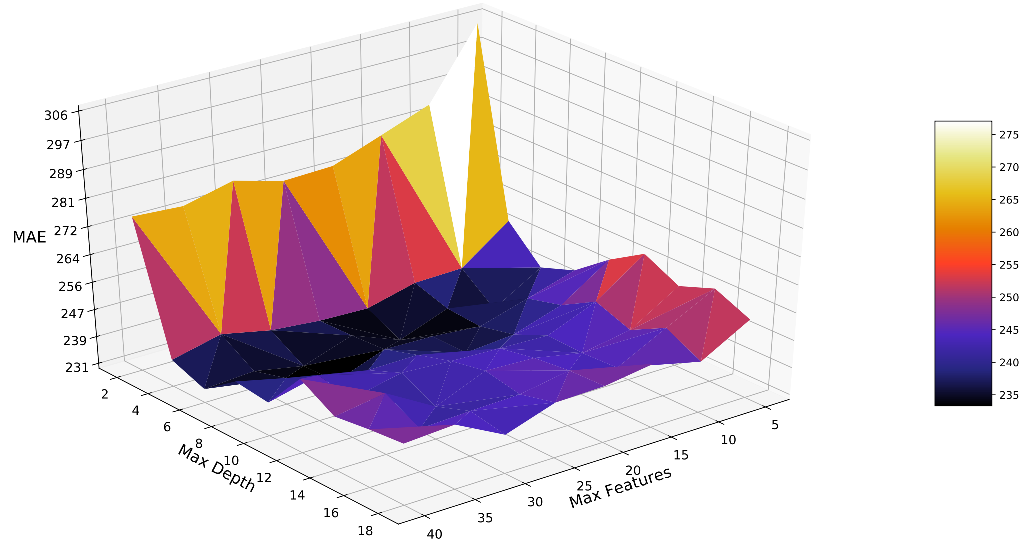

Figure 2: Gradient Boosted Trees: MAE v/s hyperparameters

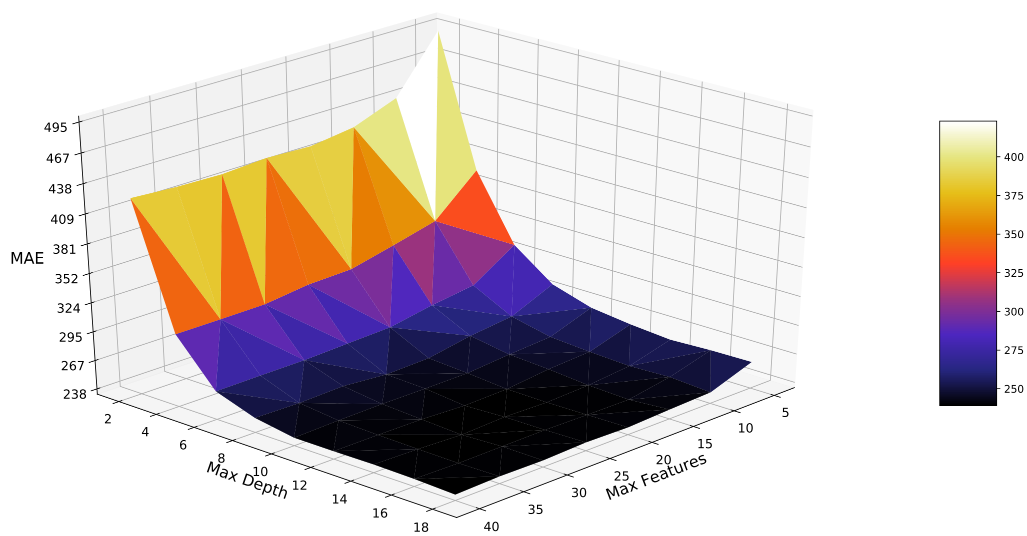

Figure 3: Random Forest: MAE v/s hyperparameters

In order to gauge the effectiveness of feature selection for our problem, we test the performance of our model, with and without reduced features. We expect better feature independence in the reduced dataset, and thus a greater reduction in variance in our ensemble based methods. However, this gain may be offset by arguably losing a few valid valid features as well. On testing, we observed that the MAE showed no significant deviation, while gained a significant speedup due to reducing features from 110 to 67.

Hyperparameter optimization for GBTs and RFs was done by exhaustive grid search. An iterative search over the parameter space yielded results as seen in figures (2) and (3) above:

Feature importance values were obtained for RFs and GBTs. OLS coefficents were scaled down by a factor of their absolute sum, thus making them directly comparable to the ensemble's feature importances vector. Computing Feature importances for the original features in SVMs and ANNs, and was not attempted in this project due to the greater difficultly framing experiments to obtain them.

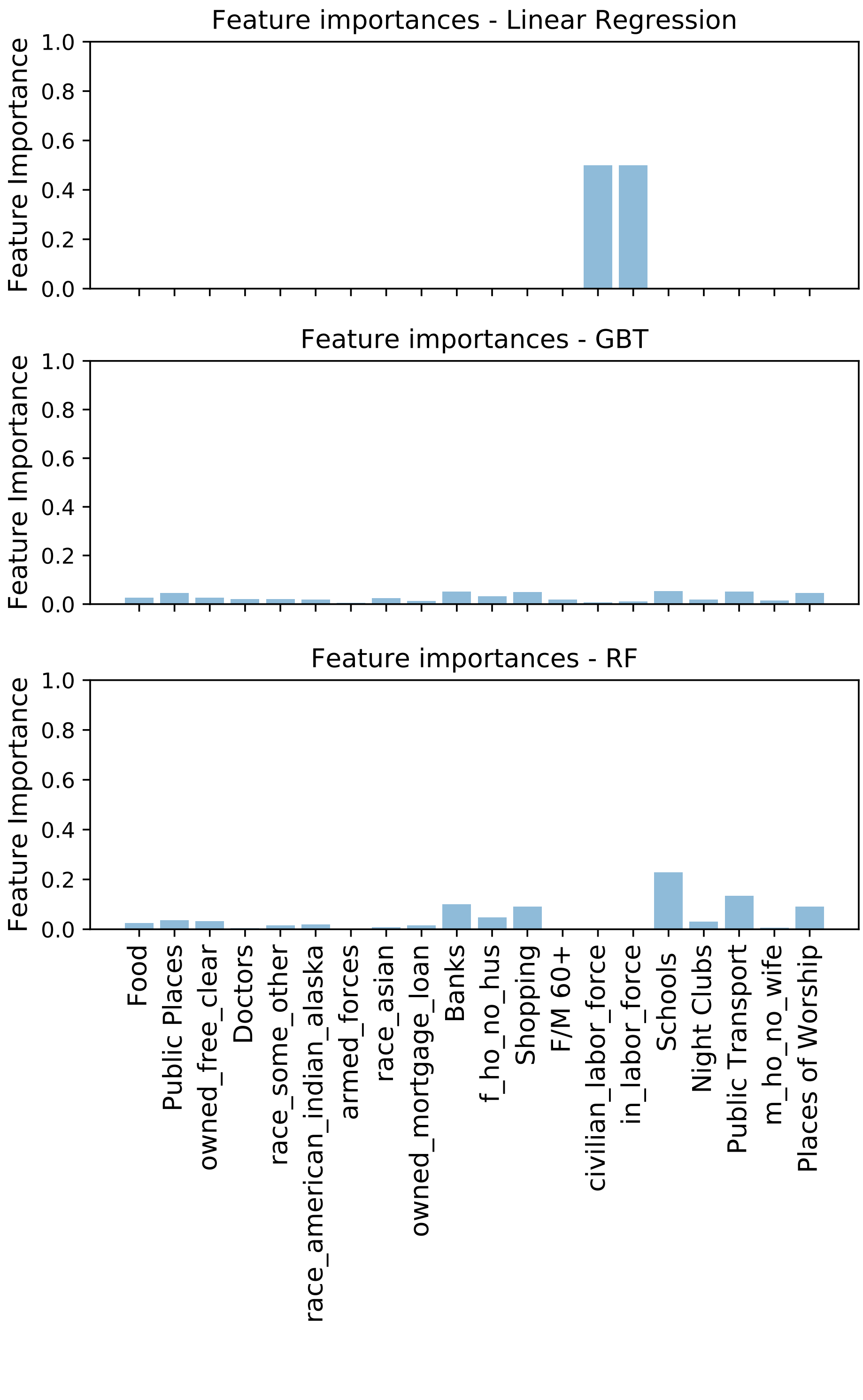

Figure 4: Feature importances for Linear Regression, GBT and RF

We have taken the top 3 features from Linear Regression and top 15 from GBT and plotted them above. The reason we only chose the top 3 features in LR is because LR reported high feature importance on only 2 features and the rest were almost 0. In total, we got 20 unique features out of these. LR places high importance on the number of people in the labor force and the number of people in the civilian labor force. GBT and RF much less importance on these two features. GBT and RF put high importances on the number of Food places, Schools, Public Transport (Train station, Bus station etc.), Banks, Public places and Shopping places. This is unsurprising as Broidy *et al.* [15] showed that there is a high correlation between schools and crimes. GBT and RF also place high importance on nearby Night clubs which is contrary to the study done by Linz *et al.* [16]. A very high relative importance is placed on `f_ho_no_hus` (Female householder with no husband). The feature importances also suggest that there is a very high correlation between the number of Banks and crime incidents. This is corroborated by the findings of Garmaise *et al.* [17]. They found that there is a negative correlation between the number of banks in an area and the crime rate. In fact, they even went further and showed how less bank competition in area leads to worse credit terms, less investment and low property prices and how less bank competition directly caused an increase in crime.

5.1 Inferences

The prediction error (MAE) obtained was comparable to the bin size chosen in cases where the problem was framed as a classification problem with 10 class labels, which is higher resolution than most past research endeavors in the area which have stratified risk zones into 3-5 class labels for risk values. We thus conclude, that our model is a competent alternative to existing models. In addition, given that we utilize a much more unique feature set than most previous studies, the model may serve as an excellent addition to preexisting ensembles of crime predictors, due to its potential independence from other models in the ensemble.

There was a surprisingly low emphasis on race in both of the tested regressors. Thus, our results do not support commonly held preconceptions that lead to demographic profiling. However, given the narrow scope of our experiments, we do not have sufficient knowledge to refute these preconceptions either.

There was a very strong positive correlation that was drawn between highly active areas with a lot of infrastructure, supported by both RF and GBT models.

5.2 Caution

The feature importances obtained from all the methods, establish merely correlation and not causation. Often a highly weighted feature might be a symptom of the real cause that the model does not capture.

6 Discussion and conclusions

In this project, we proposed a novel approach to predict crime for neighborhoods in a city from multiple sources, in particular nearby landmarks, census and demographic data. We were able to achieve results that were significantly better than a baseline linear regressor. Many prior studies have tried to predict crime based on demographic and socio-economic data, but there has been no prior work on prediction of crime based on landmarks data. We were able to successfully show that local infrastructure data plays a more important role in predicting crime incidents. We also showed the importance of different features on crime rate. We were able to confirm the results of prior studies related to correlation between crime, and schools and banks[15, 17].

6.1 Future work:

In our project, we are predicting the crime rate of a neighborhood given information about landmark such as the number of restaurants, alcohol shops, public places etc and the socio-economic conditions and the demographic information which is obtained from the census data. While this is an interesting and useful problem to solve, there are several things we could add to improve our model. In addition to predicting the crime rate or the number of crimes over a 2 year period which is a regression task, in future, we could also develop a classification model that could predict the highest category of crime that were likely to occur in a given neighborhood. The crime data set that we are using mentions the various types of crimes that occur in each neighborhood and thus one would not need any additional information or data set to accomplish this task. As outlined in the 2, it has been shown several times that adding contextual social media information such as Twitter data has a huge impact on the performance of crime prediction. We could incorporate similar data from social-media and it might improve our results.

For this project, we have neglected several time related factors such as the time of the day a crime has occurred, the day of the week a crime has occurred. In future, one could use such information and extend our model to predict the time of a day a crime was most likely to occur. This could be a useful prediction and as law enforcement agencies could be prepared for such a deed in advance.

We hope that this project will help decision makers and city officials allocate resources in areas that have undergone significant changes in infrastructure and demographics. We also hope that future studies on crime prediction concentrate on local infrastructure data as we have shown that it has high importance in crime prediction. Finally, we hope that our project has some contribution to the eradication of the perennial problem of crime.

References

[1] Isaac Ehrlich. On the relation between education and crime. In *Education, income, and human behavior*, pages 313–338. NBER, 1975.

[2] John Braithwaite. *Crime, shame and reintegration*. Cambridge University Press, 1989.

[3] E Britt Patterson. Poverty, income inequality, and community crime rates. *Criminology*, 29(4):755–776, 1991.

[4] Halvor Mehlum, Karl Moene, and Ragnar Torvik. Crime induced poverty traps. *Journal of Development Economics*, 77(2):325–340, 2005.

[5] Stephen Machin, Olivier Marie, and Sunčica Vujić. The crime reducing effect of education. *The Economic Journal*, 121(552):463–484, 2011.

[6] Zach Friend. Predictive policing: Using technology to reduce crime. *FBI Law Enforcement Bulletin*, 82(4), 2013.

[7] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pages 427–434. ACM, 2014.

[8] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1-2):4–28, 2008.

[9] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer, 2013.

[10] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding hotspots. 2005.

[11] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 231–238. Springer, 2012.

[12] Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[13] US Census Bureau. Developers, Sep 2016. Decennial Census 2010.

[14] US Census Bureau. Developers, Sep 2016. American Community Survey 1-Year Data (2011-2015).

[15] Lisa Broidy, Dale Willits, and Kristine Denman. Schools and neighborhood crime. *Washington, DC: Justice Research Statistics Association*, 2009.

[16] Daniel Linz, Bryant Paul, Kenneth C Land, Jay R Williams, and Michael E Ezell. An examination of the assumption that adult businesses are associated with crime in surrounding areas: A secondary effects study in charlotte, north carolina. *Law & Society Review*, 38(1):69–104, 2004.

[17] Mark J Garmaise and Tobias J Moskowitz. More banks, less crime? the real and social effects of bank competition. 2004.