Patrick's Writeup for Team: Data Squad

Predicting Homicide Rates From Socio-Economic Factors

Introduction:

Crime is a bane of human society but it is ubiquitous in all cultures and countries. Having a low crime rate in society is extremely important for the well-being of its people and the progress of the country. Therefore, lowering crime rate is a highly desirable objective. In order to achieve our objective of lowering crime, we need to be able to predict crime. **The objective of our project is to utilize indicators of socio-economic progress to predict crime. This predictive capability then provides us the means to lower crime by controlling and directing these socio-economic factors in desirable directions**. **The predictive capability also provides us informed advice on the most appropriate distribution of economic resources amongst the various sectors that would have the maximum impact on lowering crime**. Socio-economic indicators are routinely measured for many other purposes such as part of economic planning, for quantifying targets to be achieved and as measures of effectiveness of existing policies. **So, this data already exists**. Because machine learning algorithms have the capability to reliably capture associations, we can squeeze even more utility from this data (than its original intended purpose) by using it as input into machine learning algorithms and discovering associations between hitherto unrelated variables. Because collection of reliable socio-economic data is a very expensive activity, the ability to use machine learning algorithms to squeeze additional utility from already existing data is highly desirable. Crime is a complex and multi-faceted phenomenon. It can be explored at multiple levels. For example, we can ask questions such as: Are the factors that we can use to predict crime economic in nature, social in nature, geographical or a combination thereof? Is there a historical context to crime? In addition, crime can be investigated (for the sake of predicting crime) at different geographical levels (country, city, neighborhood), intention of crime (intentional vs. accidental) as well as on the violent nature of crime (theft vs. sexual assault vs. homicide). It is quite evident that crime is a multidimensional problem and is an enormous area of research by itself. Therefore, for our goal of a class project-which is to be finished within a limited time-we would explore only certain aspects that can help us predict crime. However, the methods, models and techniques that we would introduce in this analysis are quite general in nature and are applicable for more complex analysis of crime prediction as well as other predictive analytics problems.

Problem Statement:

As mentioned above, crime is a complex phenomenon. So, we will like to first focus our query on the following:

We will like to predict **homicide rates** at **country level** based on the **economic indices of the country's prosperity**.

This question is significant because even in our everyday experience, in our wonderful city of Chicago, we have some inkling that certain neighborhoods of Chicago are more prone to crime than others. A formal, data-based study to predict homicide rates at country level, based on the economic indices of the country is, therefore, important. **It is worthwhile to mention here that although this project will use country as a unit of analysis, the general nature of our analysis allows the same modeling techniques to be utilized for smaller geographical units such as cities and counties, albeit with some modifications**. Because there are large number of predictors of economic prosperity (but the number of countries and the countries for which data is available are limited-and we have to do a training and test split of the dataset as well), we will have to limit the feature space to 5-7 predictors for our **primary data set** so as to not run into the problem of "curse of dimensionality". Then, **as a first secondary dataset**, we will be augmenting our primary dataset with 3-4 additional predictors on woman empowerment in the country so as to compare the difference between the performance of our predictive model based on the primary dataset and when the primary dataset is augmented with first secondary dataset. As a **second secondary dataset**, we will be using 3-4 additional predictors on the educational development in the country to check for additional improvements in our model when the predictors from the second secondary dataset are included in the model. It is noted that additional features from secondary datasets increase the dimensionality of the feature space.

As economic planners for the country, we will like to know the answer to the following question:

Which factors or combinations of factors are better predictors of homicide rate?

The answer to the above question will help the economic planners to answer the following questions:

Of the three sectors that we can invest in, which investments or a combination of investments have the biggest impact on lowering homicide rates?

In the context of lowering homicide rates, should we focus more on improving economic prosperity, on woman empowerment or on the educational development in the country or combine some of these factors to get the maximum impact for lowering homicide rate?

Data Sources:

We will be using the data from the following source:

https://ourworldindata.org/countries

This data source contains data on large number of socio-economic indicators for almost all the countries in the world under various categories including economic, gender and education. For example, for countries such as US and China more than 700 indicators have been provided. As noted above, from the viewpoint of data, the bottleneck in our analysis is not the limited

number of socio-economic indicators. It is the number of countries compared to the number of possible socio-economic indicators so that we do not run into the "curse of dimensionality" problem. This data source was chosen because the data is reliable, open source and the sources of data have been cited on the website. Another reason for selecting this data source is the quantitative nature of the predictors and the target variable-the economic and social predictors that we will be using in our model have an **interpretable, physical basis** and make business sense.

Project Outline:

The first step of our analysis will be to gather and compile data from the above data source and arrange it in the form of tidy data so that it is amenable for analysis. In tidy data, each observation has its own row, each attribute has its own column and each value has its own cell. In gathering data, we will be mindful of the problem of "curse of dimensionality", because the number of possible predictors far exceeds the number of countries.

When the data has been arranged into tidy data format, we would explore the integrity of the data and answer some basic question: Are there missing values in the data set? If there are missing values in the dataset, how should we take care of them? Should we impute them or delete the corresponding rows? If we decide to impute them, what methods should we use to impute these missing data values?

After we have done some basic sanity checks on the data, we would like to do exploratory data analysis to better understand the data, get a feel for the variables in the dataset and to perform further sanity checks on the data. For this, we will use both numerical and visualization techniques to get a better sense of the variables. We will use appropriate measures of central tendency and spread to better understand the individual variables as well as look for issues in the dataset. We will use boxplots and histograms to visualize our data, discover outliers and check for any inconsistencies in the dataset. In addition to yielding important information into the nature of variables, we would be able to perform additional checks on the sanity of data by answering questions such as: Are there any erroneous values in the dataset? Do the data values make business sense? Then, we like to explore relationships between variables to understand how variables correlate to one another and whether these correlations are consistent with our business sense. The requirements that the interrelationships between variables make business sense provides an additional sanity check on the dataset. These relationships will be explored using scatterplot matrix and cross tabulations for numeric and categorical variables respectively.

After we have gathered the data, arranged it in the form of tidy data, done a thorough investigation of the dataset to understand the various variables in the dataset and their interrelationships, checked it for consistency and sanity, we may need to transform/standardize some variables because some of the predictive algorithms that we will be using in this project are sensitive to the scale of the variable. If some of the predictors in the dataset are categorical

variables, they need to be suitably encoded so that they can be used as input into regression algorithms that we will be using to build a model.

After standardization, we will like to use regression modeling techniques mentioned in the next section to construct models. To control the bias-variance trade off and because the number of records in our dataset are limited, we will use cross-validation to train the hyperparameters of our models and to reliably predict the performance of our models. The final model, with chosen values of hyperparameters, will be trained on the entire dataset so as to ensure the most optimum values of parameters in the final model.

Predictive Modeling Techniques:

In choosing a predictive modeling technique for analysis, there is a tradeoff between predictive power and interpretability of the model. A complex modeling technique can capture more intricate associations between the variables but is in general also less interpretable. There is also an inherent bias-variance tradeoff and one is forced to accept a certain optimum combination of bias and variance. Keeping these considerations in mind, and because our target variable is numeric in nature, we will be using the following statistical modeling techniques in this project:

1. Artificial Neural Networks (ANN): ANN are capable of capturing extremely nonlinear relationships between the variables but they are in general less interpretable compared to methods that relate variables in an equation
2. Linear Regression: Linear regression is more interpretable than ANN but depending upon the problem, may not be able to fully capture a highly complex relationship.
3. K-Nearest Neighbor (KNN): KNN has a simple interpretation that the target value for the input observation is the mean of the target values of its neighbors.
4. Lasso: Lasso is a regularization technique that is used to reduce overfitting. Lasso can also act as a variable selection technique because it can result in the coefficients of certain variables to be set to zero.
5. In addition, we will be presenting some results using TensorFlow, using Keras as an interface to it. We understand that using TensorFlow for this problem will be an overkill because TensorFlow was designed to handle several thousands of predictors. However, it is anticipated that the knowledgebase and workflows that we will develop in using TensorFlow will be an extremely useful learning experience for future projects.

Validation and Evaluation Techniques:

We will use 10-fold cross validation to evaluate our models. This is an industry standard.

KPIs:

The performance of regression models is measure by two factors:

1. Root mean square error (RMSE).
2. Correlation between the predicted and actual value of target variable.

We will be using both these measures to quantify the performance of our models.

Success Metrics:

We will like to have our RMSE to be low and desire a high correlation between the predicted and actual values of the target variable. The actual numerical values for the above mentioned KPIs, which we will get for our dataset would only be known after the model has been constructed. As mentioned earlier, there is a tradeoff between interpretability and predictive power of the model. That is why we have chosen a diverse spectrum of techniques from ANN to Linear Regression because we canny say a priori as to how complex the relationship between the predictors and the target variable is. In addition, Lasso may eliminate certain variable from the analysis providing a simpler, more interpretable model which may have performance comparable to ANN. The conclusion is: while we do have a general direction for the values of KPIs, we will need to complete the analysis first to give precise values of KPIs and the actual numbers for KPIs also depend on the dataset at hand.

Deliverables:

The deliverable will be a predictive model and also a report based on conclusions from the predictive model that can be used to predict homicide rates based on socio-economic factors. The report would contain information on the relationships between the socio-economic factors and how they affect the homicide rate, which would allow the economic planners to build strategies for investment in various socio-economic targets so as to have maximum impact on homicide rates. A summary of technical aspects of the techniques used for modeling will also be provided. A executive summary of our experience with using TensorFlow for this project will also be provided. As mentioned earlier, TensorFlow may be an overkill for this project but the TensorFlow workflows and knowledgebase that we will develop for this project will be usable for future projects as well.