**Project Problem and Hypothesis**

*What’s the project about? What problem are you solving?*

This project is about predicting crime resolutions in San Francisco. It is trying to predict whether or not a particular type of crime will be resolved or not, given its crime category, the day of the week, the district the crime was committed in, and its location.

*Where does this seem to reside as a machine learning problem?*

This project will try to predict a binary outcome: whether or not a reported crime was resolved or not.

*What kind of impact do you think it could have?*

The impacts of this project could be twofold: one positive, but one negative as well. If authorities are able to better understand what kinds of crimes are going unresolved, they can better allocate resources to try to cut down on those that go unresolved. The other side of that coin is that if criminals had a better understanding of what kind of crimes go unresolved, they are more likely to be able to target certain regions, or times of day, or types of crime, knowing what they are more likely to get away with.

*What do you think will have the most impact in predicting the value you are interested in solving for?*

I suspect that the type of crime will be the strongest predictor of whether or not that crime will be resolved or not. Certain crimes are almost certainly more likely to be resolved than others. For example, an assailant may be more likely to be caught than a white collar criminal (or vice versa).

**Datasets**

*Description of data set available:*

Dates – timestamp of the crime incident

Category – category of the crime incident

Descript – detailed description of the crime incident

DayOfWeek – the day of the week

PdDistrict – name of the Police Department District

Resolution – how the crime incident was resolved (this is the outcome variable)

Address – the approximate street address of the crime incident

X – Longitude

Y – Latitude

**Domain Knowledge**

*What experience do you already have around this area?*

I do not have any experience analyzing or understanding criminal activity, but I am a resident of the city and have a general understanding of the crime that happens within it.

*Does it relate or help inform the project in any way?*

Certainly. General street smarts will help in this project. Already thinking ahead to the feature engineering stage of this project, there are many things that come to mind as potential predictors of whether or not a crime will be resolved. What street did the crime occur on? How far is the location of the crime to the nearest police precinct? How far is the crime from the PdDistrict precinct? Answering these kinds of questions might go a long way to predicting the outcome of crimes in San Francisco.

*What other research efforts exist?*

Given that this dataset comes from a Kaggle competition, there is a lot in the way of public research that is available online. Even though I am attempting to answer a different question than the Kaggle competition, many people have attempted to dig into this dataset and better understand it.

**Project Concerns**

*What questions do you have about your project? What are you not sure you quite understand yet?*

There is a lot of text data involved here, and that is something I don’t have a lot of experience analyzing. Granted, much of it is in the way of a categorical factor variable (crime category, day of week, police district, etc.), I think there is probably a lot of information to be gained from the crime description and address fields that are not simple categorical variables. Also, doing some introductory exploration of the dataset, I noticed that the reporting of crimes seems to change somewhat over time, and I think that could be problematic. Additionally, there seem to be an awful lot of crimes that are reported at midnight, more than one would probably expect. After doing some research, it seems that this is just an element of the police crime report process that oftentimes crimes are reported at midnight for bureaucratic reasons, not because that’s when the crime was actually committed. This makes me doubt the entire date and time field, and although I think it will be important, it concerns me a bit to use it.

*What are the assumptions and caveats to the problem?*

I am assuming that every crime that is committed in the city is reported, but obviously that is not the case. There are crimes that simply may be more likely to get away with than others, based on any of the factors that I actually have in my dataset, that I will never know simply because I don’t have data on those crimes as they were never reported.

*What are the risks to the project?*

Certainly, any of this data could be wrong. I touched on it above: bureaucratic reporting standards could inhibit the success of this project. Additionally, this dataset goes back 12 years, and reporting standards could easily, if not probably, change in that period of time, so that the dataset I am working with is more dynamic than I would like it to be.

**Outcomes**

*What do you expect the output to look like?*

I expect to see a probability for any given crime of that crime being resolved or not.

*What does your target audience expect the output to look like?*

My target audience may want to see the output in the same way I do; i.e, plug in a crime, see a probability of that crime being resolved. But they might also like to see other output, like an interactive map that allows you to select a crime and a time of day, for instance, and see a heatmap of the likelihood of someone getting away with that crime. I don’t know if that is something that I can actually produce, but it would be awesome if so.

*What gain do you expect from your most important feature on its own?*

As I mentioned above, I think the crime category will be a very significant driver in predicting whether or not a crime will be resolved, so suffice it to say that I expect a lot of gain from that feature alone. I also think the crime’s proximity to certain things in the city could be a driver. The police precinct for one, but there are a lot of other things to consider. What kind of street was the crime committed on? Residential or commercial? High income or low income? I do expect a fair amount of gain from these features, if I can acquire them.

*How complicated does your model have to be?*

I don’t anticipate the model itself having to be all that complicated. What I expect to be complicated is the feature engineering portion of the project. Anybody can throw these variables into a logistic regression and maybe come up with a half-decent model, but where I can really take my project to the next level is in coming up with creative features to include.

*How successful does your project have to be in order to be considered a success?*

It is really hard to say without having run any test models at this point; I think only then will I have a good benchmark for success. Intuitively, I’d certainly like to do better than a coin flip, or better than the fraction of all crimes that are resolved. But also, intuitively, I don’t expect to be getting 95% of these right. There is a lot of noise in this dataset, and a lot of factors that are simply unavailable (did the victim report the crime promptly? Was there already an officer nearby?).

*What will you do if the project is a bust?*

This project is meant to be a learning experience for me. If I put forth my best effort and the project does not succeed, I will have learned just as much as if I put forth my best effort and do succeed. It would simply mean that I would have to reevaluate my process and reevaluate my question and think about how I could make it more successful in the future.