Report

Preface: Please read through Analysis.ipynb first

After analyzing many parts of the gun deaths dataset, I came to a few conclusions on how to build a model that would maximize performance. As seen in my analysis, the month at first glance seemed to tell us some stuff since in February there was lower gun death rates amongst all categories of intent; as well has slightly higher rates in July. With high confidence, I believe the February numbers are since it is a shorter month. The July part was interesting though. Even though July was slightly higher than any other month, I thought it best to just remove the entire feature from the dataset. In the future, there may be a way to scale February out so that we could maybe use July to help our model; though I still do not think it would affect prediction accuracy much. After looking at the Hispanic feature, over 90% of the codes were indicated as non-Hispanic, so this feature was useless. Especially given that most of the deaths were seen to affect whites and blacks. So, Hispanic column was removed. As for places, this could have been a good deciding factor, but the distribution was not specific enough in a single area. The only thing the model could have maybe gotten from this data is that if the incident occurred at home, there was a higher probability that it may be a suicide. In the end, I created a few different machine learning classification models. I went in thinking that SVC would dominate all. But in fact Logistic Regression performed better than the others. After my analysis, I found that removing the fields that I did, did increase prediction accuracy; it outperformed all other models that I saw on the Kaggle Kernals. Though, it only improved upon the control predictor (nothing removed) by a maximum of about 5%. This is obviously better than nothing so it worked out well.

In the future, I would want to create a different kind of model. One where it looks down more than one dimension, for example: if Age and Sex = AS, and if Location and Race = LR, then we build a model that actually weights features such as AS and LR. But I believe that goes into deep learning and neural networks. I think a model like that could be far more accurate at predicting given that my analysis was similar.