Summary

In goal of this project was predicting arrests based on crime information given by the Chicago police department. As I have mentioned above, only 25% of crimes end up with arrests, which makes a challenge due to unbalanced dataset (which is partly resolved by stratify). I addition, Since a significant amount of data is categorical and not ordered - the usage of decision trees is limited. Finally, while some features that describe the type of the crime had very high correlation with the target, the rest had a very low correlation. I suspect that this leads to the fact that the models put extremely high emphasis on the important features, and thus the rest of the features contribution to the classification is low. While the accuracy we have seen is not bad at all (~89%), in order to get more improvement, one might need to use more complex models to get the best out of the features that had low correlation with the target.