MDML_Assignment 4
Sameen Reza
Andrew Chase
Rami Tello

A3.1

The fewest number of crimes are committed between 5:00 am and 7:30 am. From 7:30 am to 8:00 pm, the volume increases in a linear fashion. Around 8:00-9:00 pm, however, the volume of crimes committed spikes. This spike then lasts for four hours, until 12:00 a.m, and then begins to settle down. Between 12:00 am and 2:30, the volume of crimes committed is halved as compared to the volume of crimes committed between 8:00 pm and 12:00 am. The primary takeaway from this graph is that most crime in Boston occurs at night: specifically between the hours of 8:00 pm and 12:00 am.

A3.2

This graph shows an uptick in almost all of the top 5 crimes between the hours of 20:00 (8:00 pm) and 22:00 (10pm) in Boston. The exception to this trend is illegal gun possession, which peaks around 12:30 (12:30 pm). Perhaps the peak occurs at this time because it is easier to see at this hour as opposed to looking for weapons when it is dark outside. This graph lets us view the relative frequency of the top five crimes. It is curious that shooting spikes between the hours of 20:00 (8:00 pm) and 24:00 (12:00 am), but illegal gun possession does not. Perhaps shooters (who are carrying guns) are not charged with illegal gun possession if a shooting charge is levied.Between 5:00am and 12:00pm most crimes remain at a stagnant low.

A3.3

This graph shows that Dorchester exhibits a far higher volume of crimes than downtown Boston. Dorchester tends follows the same pattern exhibited in A3.1 and A3.2, with most of its crimes occuring at night. On the other hand, downtown Boston experiences most of its crimes between the hours of 2 pm and 5 pm albeit at a far lower volume than Dorchester.

B2.2

Precinct, stopped.bc.bulge, stopped.bc.object, max AUC = 76.86%

B2.3

New_AUC = 75.57%

B2.5

Model B2.3 was trained on a larger set of data 80% so it seems to be a better model. The dotted line would give a better estimate of the expected AUC of this model on unseen data.

B3.4

Sqf_pre_test: 0.816
Sqf_2015: 0.750

B3.5

Write a paragraph answering the following questions:

i. Why do you think the AUC on sqf_pre_test is noticeably higher than the AUC on sqf_2015?

Sqf pre test is 50% of the data and the model was trained on the other 50%. We think the model is overfitted given the size of the data on which it was trained and that this is the reason that it gives a high performance on that data. However, when used this model on unseen data of 2015, its accuracy is not very high.

ii. If you were planning to use this model to guide how officers make stops in the future (e.g., by having officers use the model to compute the probability that an individual suspected of criminal possession of a weapon will have a weapon, and then only making a stop if the model-estimated probability is sufficiently high), would the AUC on sqf_pre_test or sqf_2015 be a better estimate of performance on unseen data?

The sqf_2015 auc would be a better estimate of performance on unseen data

iii. More generally, when evaluating a model using a simple training/validation split approach, should you always do the split by shuffling and splitting randomly?

We think that although shuffling and splitting is an easy to use and practically feasible approach, the data needs to be looked into in detail to see if this approach makes sense. For instance, it may not be an appropriate approach for time series data or where certain events are known to have happened in a certain year.

QB4.2

In your writeup, report what you see in this plot, and why this might occur.

The model has been trained on data from 2008 and is being to predict for data from 2009 onwards. The model performs consistently for years uptil 2013 but after 2013 the model performance gets reduced. I think this is because the data set became different with lesser number of observations and new variables were introduced after the law making such checks unconstitutional was passed. This newer data is not representative of the data that the model is trained on. Which is the cause of the performance degradation.