Medium article: [*Predictive Modeling: Picking the Best Model*](https://towardsdatascience.com/predictive-modeling-picking-the-best-model-69ad407e1ee7)

* Logistic regression, random forest, k-nearest neighbors, XGBoost
* Used pandas, numpy, scikit learn

Wikipedia on statistical binary classification models

* SVM
* Decision trees and random forests (What’s the difference between CART and random forests?)
* Bayesian networks
* Neural networks
* Would it be helpful to read up on these, (also in ISLR), and choose a few to test out? Read about and try out at the same time? What is the process of trying out these different models…

Medium article: [*Traffic accident predictions*](https://towardsdatascience.com/live-prediction-of-traffic-accident-risks-using-machine-learning-and-google-maps-d2eeffb9389e)

* Heat maps
* Negative sampling procedure based on Yuan… for every traffic accident in a location, generate 3 non-traffic accidents with random time/date information…
* Dark Sky API, which has a corresponding R package, can give the weather data by the half hour. Cloud cover, precipitation rate, nearest storm distance, wind…
* DBSCAN algorithm to perform this clustering: cross section of 25 meters that has more than 14 accidents in 2 years. Identifies accident-prone places

(Not relevant) Undergraduate thesis: [*Predicting hit rates in Illinois, statewide*](https://economics.uoregon.edu/wp-content/uploads/sites/4/2020/01/Jordan-Hamada-and-Linnet-Sim-Predictions-Contraband-Econ-419-Thesis-2018.pdf)

* What is “OLS Algorithm predictions” …
* I could… make something better than this. Hit rates have the benefit of not being a rare event (contraband is recovered in about 20% of stops)

A picture containing text

Description automatically generated

Reverse Geo-coding: [revgeo](https://cran.r-project.org/web/packages/revgeo/revgeo.pdf) package in R!!! Can take coordinates and give county!

Article I can’t access: **Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study**

* Compare SVM and Multilayer Perceptron
* Use Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance

Medium article: [Comparison of Machine Learning Classification Models for Credit Card Default Data](https://medium.com/@vijaya.beeravalli/comparison-of-machine-learning-classification-models-for-credit-card-default-data-c3cf805c9a5a)

* Looks at 10 different ML methods for credit card default, using R to implement
* Identifies strengths/weaknesses of method (haven’t read it fully, though)

Bookdown: Statistical Learning with R textbook

* [Caret package tutorial](https://daviddalpiaz.github.io/r4sl/the-caret-package.html), with cross validation

Questions @ self

* What is Lasso and what is Boost?
* Identify “clusters” of stops to see where there are hot spots, find the major highways through those clusters, then limit the stop/search behavior to be only those on the major highway? Because major highway stops are probably all low-discretion.
* Include month, year, day of the week as variables? Seasonal variations on stop/search based on budgets???

Questions

1. Linear / quadratic discriminant analysis is not necessarily applicable, right?
2. Unsupervised learning on traffic stop data?
3. (Imputing) missing values? / Chanin
   1. Missing values in general (race, age, …)
   2. Missing values for time and therefore sunset/sunrise

Future tasks (to self study)

* Narrow down the models to try out
* Goal: learn Python through this? Hm…
* Learn/write about
  + Cross validation techniques
  + Logistic regression
  + Support vector machines

August 10th Meeting Agenda

1. (the difficulty of) Defining the scope of the project
   1. Knox paper: underlying bias of the data, that race itself is a causal factor of being stopped. “post-treatment bias”
      1. P(search) = P(searched | stopped)\*P(stopped | race)\*P(race)
   2. Yuan: negative sampling
   3. Missing data
   4. High/low discretionary stops – find it through clustering to find highways?
   5. Spatial heterogeneity and changes through time. Each department has different practices, but also each unit has different practices. It would be better, for example, to compare highway patrols across the nation.
2. More specific experimental design notes
   1. VOD … is it even necessary to control for time of day? Maybe, bin the stops into like 3 hour-ish periods? And have a dark/light/ambiguous categorical variable? Right now, subsetting for traffic stops at a particular time, losing a lot of data points.
3. Difficult to find comparison of binary classification models regarding traffic stop/policing. Found a few papers on accident predictions, mainly seem to be on epidemiology/disease
4. Linear discriminant analysis?