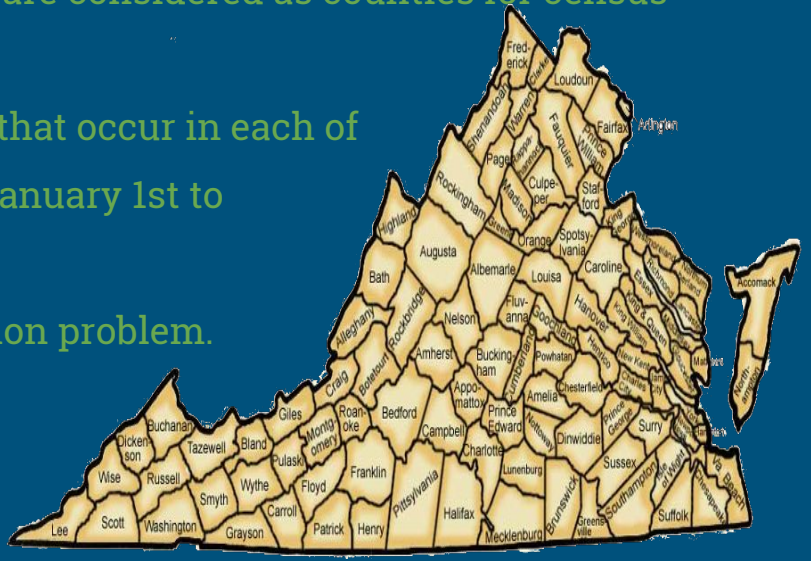
Predicting car accident related fatalities in Virginia

Daniele Correa

Question: Can we predict the number of fatalities from car accidents that will occur in a Virginia county in a calendar year?

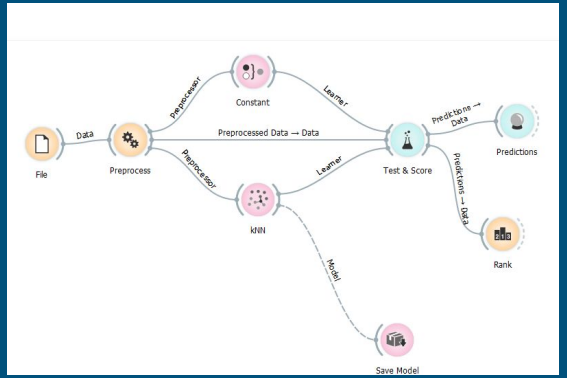
- Using specifically data from the 2016-17 calendar years.
- Included the official 95 Virginia counties.
- Also included the 38 independent cities that are considered as counties for census purposes (how I obtained data)
- Outcome variable is the number of fatalities that occur in each of these 133 census considered counties from January 1st to December 31st of a calendar year.
- Ratio outcome variable makes this a regression problem.



Data Collection:

- Collecting data consisted of parsing DMV, Census and meteorological data.
- Wrote a python script to parse the DMV report data and a web scraping script to collect the meteorological data.
- Predicting variables were: Squared Area, Licensed Drivers, Population, Teenagers , 65 or older, Car Accidents, Injuries, Alcohol Related Injuries, Speed related Injuries, Unrestrained related Injuries, I-81, I-95, rain amount, snow amount, average elevation of the county.
- The total sample size was 266, as it was data from the 133 counties for 2 years.

Model Building



- I chose the K-Nearest Neighbors Algorithm to build my model.
- I chose it because it ran very quickly and got even better results than other algorithms that took much longer to run.
- Seems to take into account relationships that are not obvious, giving better results.
- I ran this model with 10 neighbors using Manhattan distance (city-block) on standardized data centered by mean and scaled by standard deviation.

Results

- The error metrics I chose to prioritize on were RMSE and R^2
- **RMSE because it penalized large errors and because it was easier to interpret how far off my predictions were on average.**
- **R-Squa… …oodness of fit of the r…**

Method	MSE	RMSE	MAE	R2
Constant	43.097	6.565	4.825	-0.008
kNN	9.004	3.001	2.130	0.790

	#	nivar. Lin. Re
Unrestrained related Injuries	C	638.472
Alcohol Related Injuries	C	449.200
Car Accidents	C	435.583
Speed related Injuries	C	403.505
65 or older	C	394.822
Injuries	C	393.094
Teenagers	C	353.256
Population	C	342.351
Licensed Drivers	C	318.608
Squared Area	C	40.803
countyElevation	C	4.347
avgRain	C	0.365
avgSnow	C	0.290

Summary:

- Ideal future changes: Expand dataset to include many more years, get a possible safety rating score for the roads of each county, get possible safety rating score for the cars in the county, get texting and driving crash data, obtain the number of hospitals in each county.
- How my model can be applied: Government can implement safety measures in the county (e.g cracking down on people who don’t wear their seatbelts) in order to save lives.