Parking Ticket Report

Task:

Someone has deleted half, nearly 4 million rows, of the makes in the parking ticket dataset, and the make is an important tool for predicting information about the car for other models being trained. An algorithm needs to be developed to predict if the cars are from one of the top 25 manufacturers, and possibly predict which one.

Problems:

Our data set is limited in the features that we have to use, information like ticket number, issue date, issue time, agency, meter id, and marked time are not useful if we don't have the ticketing database to check against for more information. I originally thought maybe we could use the ticket description, but that information is just the description of the violation code. Route, latitude, longitude, and location are only useful if we can get camera data from that area. Leaving us with VIN, color, body style, and RP state plate. RP states are almost entirely California so that should only help a small amount as residents of states are more likely to own certain makes of cars. VIN numbers can help identify cars, but most of the dataset is made up of the 4 digit VINs which were not standardized across makes until 1981 with the introduction of the 17 digit VIN. The VIN data would only be useful in this dataset if we received a list of VINs with body types from the manufacturers.

Method:

After removing all the data that can not be used, we are left with color and body style with state plate added in to give a slightly better chance at predicting. Each color and body style was originally give their own binary row, with binary rows being also made for each possible combination. That method ending up using too many resources, so instead I opted to assigning categorized values to each color and body style and using Xgboost to predict if the make is in the top 25. I developed a model to predict if the make is a certain make or not, and I tested clustering both each make of cars and if they are in top 25 make or not.

Results:

With 90% of the cars being within the top 25 makes the data is quite skewed, I tested which class is best to label as 1 in classification and based on the performance I found it best to label the top 25 makes as one, with an increase in accuracy of over 10%. I also tested predicting each of the top 25 makes seeing if the model can identify which make it was, this gave me a precision of less than 50% and a recall of less than 30%. I did find that using a combination of the color and body type can help identify makes, but not fully After training the model multiple different ways my best performance was at a cutoff of 50% using top 25 makes as the positive class. I was able to get a precision of 92%, a recall of 99%, and an F1 of 95.6%.

Threshold	TruePositive	TrueNegative	FalsePositive	FalseNegative	Precision	Recall	F1
0.01	797312	0	74197	0	0.914864	1	0.955539
0.05	797312	0	74197	0	0.914864	1	0.955539
10	797308	41	74156	4	0.914906	0.999995	0.95556
0.15	797303	82	74115	9	0.914949	0.999989	0.955581
0.2	797268	247	73950	44	0.915119	0.999945	0.955653
0.25	797178	505	73692	134	0.915381	0.999832	0.955745
0.3	797070	806	73391	242	0.915687	0.999696	0.95585
0.35	796759	1492	72705	553	0.91638	0.999306	0.956048
0.4	796143	2598	71599	1169	0.917488	0.998534	0.956297
0.45	795637	3329	70868	1675	0.918214	0.997899	0.9564
0.5	794889	4301	69896	2423	0.919175	0.996961	0.956489
0.55	792007	6857	67340	5305	0.921638	0.993346	0.95615
0.6	789534	8554	65643	7778	0.92324	0.990245	0.955569
0.65	784910	11183	63014	12402	0.925684	0.984445	0.954161
0.7	777663	14487	59710	19649	0.928694	0.975356	0.951453
0.75	775567	15272	58925	21745	0.929388	0.972727	0.950564
0.8	772846	15961	58236	24466	0.929927	0.969314	0.949213
0.85	761090	18102	56095	36222	0.931356	0.95457	0.94282
0.9	698252	25851	48346	99060	0.935245	0.875758	0.904524
0.95	44441	73532	665	752871	0.985257	0.055739	0.105508

Improvements:

The best way to improve the model would be to receive a list of VINs from each make and include the body type for each VIN, this will allow us to easily identify the cars. If that is not possible then we can likely develop deep learning models to learn the pattern for the VIN of each make and use bootstrapping to predict if a car is within the top 25 makes. With the discovery of what data is useful it can be determined that if we had more information about the car, that would usually come from the VIN, we can likely solve this problem completely.