Bike and Car Accident Prediction Milestone Report

By Linnea Hartsuyker

Problem Statement

Many drivers hate sharing the road with bicyclists, but perhaps they could become allies, if road improvements for bikes made the road safer for cars as well and vice versa. In this project, I am examining what road attributes predict a greater density of car accidents and bike accidents, in

hopes of finding road qualities that could be changed to produce a lower accident density for both cars and bikes.

If no commonalities can be found, I will still determine what road attributes contribute to either type of accident, independently, information that road maintenance and design could take into account.

This project focuses on bike and car accidents in the greater Boston area.

Datasets

The data sets for this project come from:

- A database of Boston bike accidents from 2009 to 2012, constructed from police reports during this time: https://dataverse.harvard.edu/dataverse/BARI
- A database of car accident/crashes for the entire US in 2017
- The Massachusetts Department of Transportation (Mass DOT) road inventory https://geo-massdot.opendata.arcgis.com/datasets/road-inventory-2018
- The Mass DOT road inventory has a data dictionary

From these datasets, I needed to construct a single file that contains:

- Roughly similarly sized regions where bike accidents have occurred, car accidents have occured, and where no accidents have occured
- Each region should have a bike accident density score, a car accident density score, and road attributes

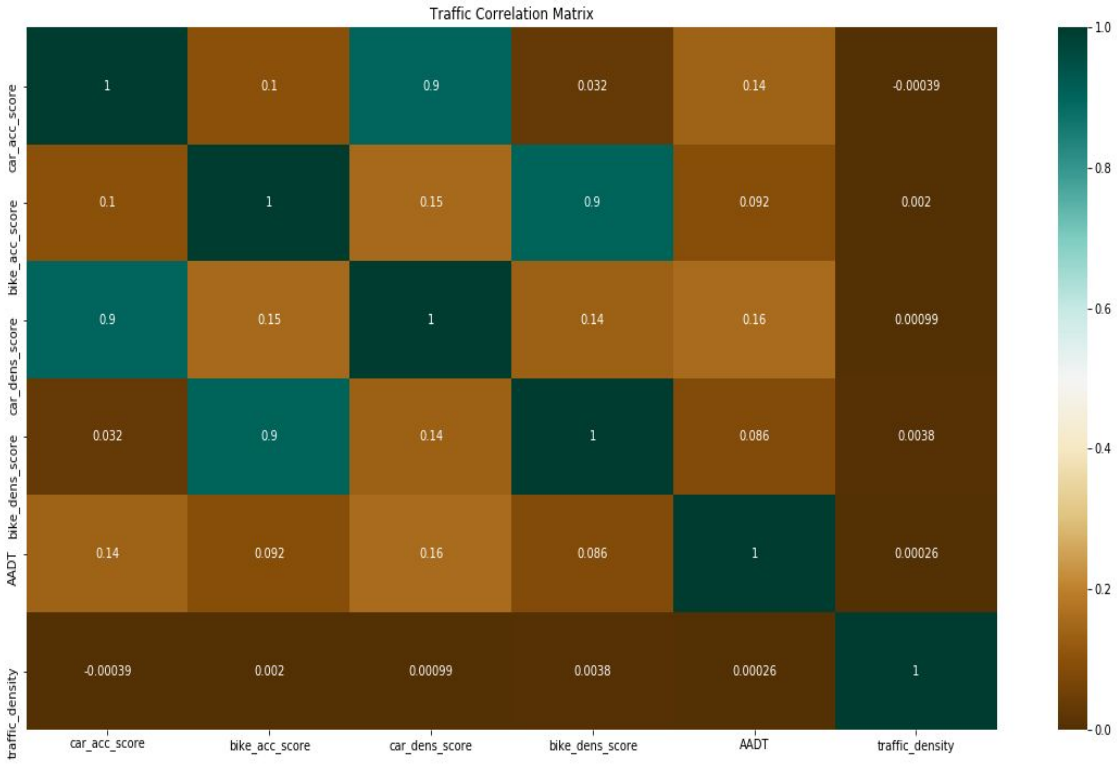
For more detail on creating this file, see the Bike and Car Accident Prediction Data Wrangling report.

Initial Findings

For code and more visualizations, see Exploratory Data Analysis and Bike and Car Accident Statistics notebooks.

Traffic

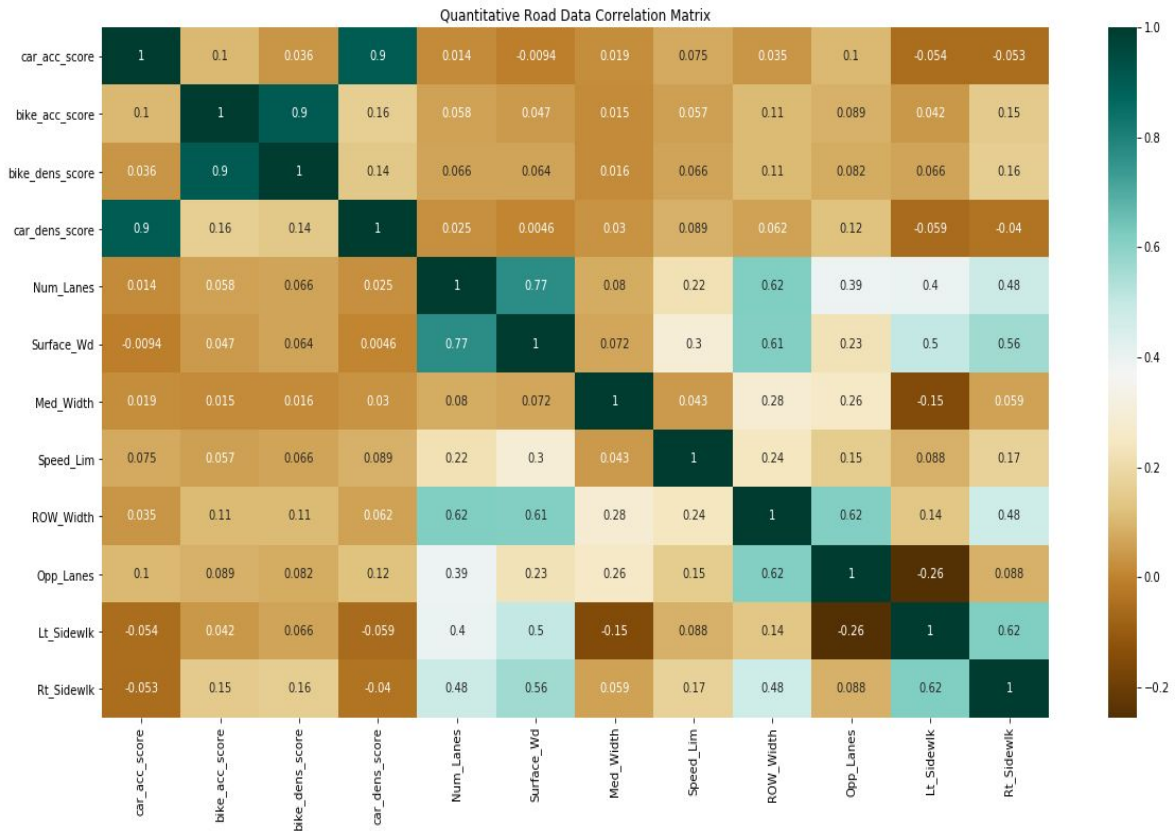
My first question was whether bike and car accidents were correlated on their own. And I was also interested in whether traffic volume was correlated with accidents. I created a data set that contained only accident and traffic data and plotted a correlation heatmap.



A correlation coefficient near 1 would mean a strong correlation, and none of the correlation coefficients are close to one, except for properties derived from one another, like accident score and accident density values.

Other Quantitative Road Attributes

Next I wanted to see if any of the numeric values had obvious correlations, with another heatmap. These values included: number of lanes, surface width, median width, speed limit, right of way width, opposite lanes, left sidewalk width, right sidewalk width, and right of way width.



This also showed no strong, direct correlations between accident scores and any quantitative values.

Digging into Qualitative Road Attributes

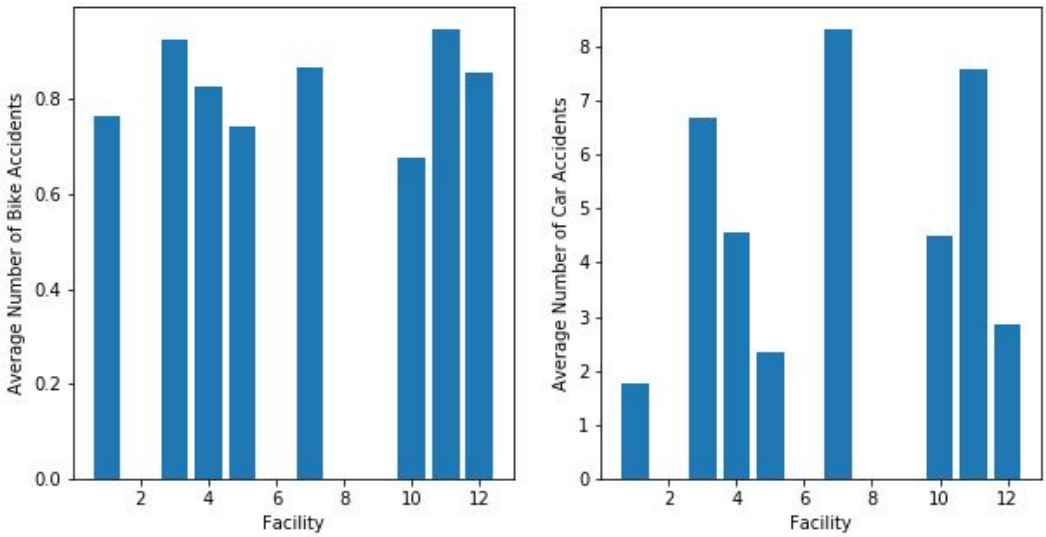
Bar charts of qualitative data against accident scores showed that several attributes appeared to be associated with high accident densities.

For reference, the average number of car accidents per area was 2.11 and the average number of bike accidents 0.75. Looking at bar charts, I compared with these values to determine “larger” average number of accidents.

For all hypothesis tests I examined the accident density scores, which are accidents per square meter, rather than average number of accidents, since this seemed like a more consistent metric.

Facility Types

Roads are designated with facility types, and the graphs showed a much larger average number of car accidents for facility types Tunnel (3), Simple ramp (7), and Simple ramp - tunnel (11), and a moderately larger average for some others. It showed a much larger average number of bike accidents for facility types Doubledeck (4), Rotary (5), Collector - Distributor (10), and Bicycle (12).

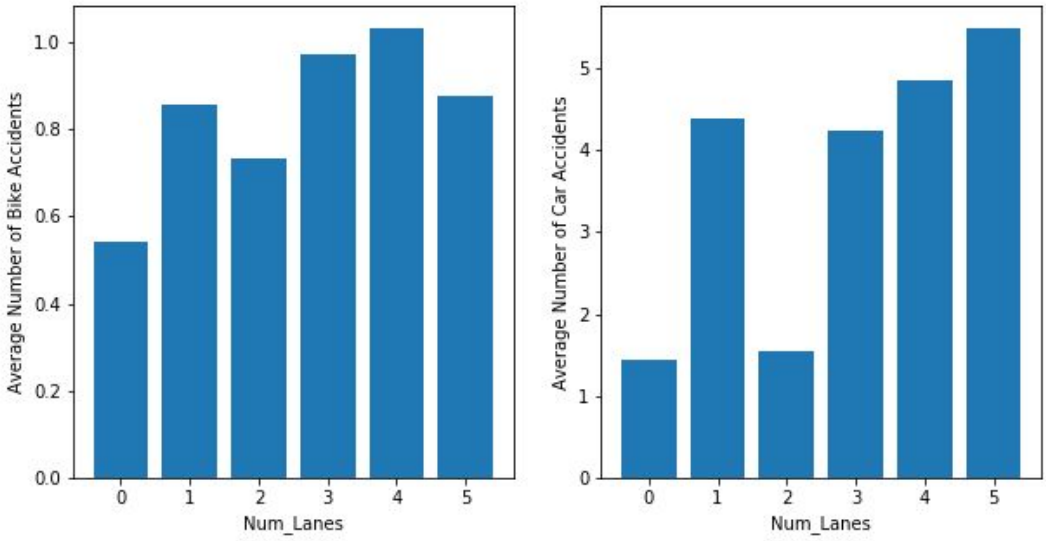


Hypothesis: Facility types 3, 4, 5, 7, 10, 11, and 12 have higher mean car accident density scores than the average across all facility types

I performed a p-test and found statistical significance to *reject the null hypothesis* that there is no difference in the mean car accident density for all of these facility types. I could only *reject the null hypothesis* bike accidents occurring in facility types 11 and 12.

Number of Lanes

Number of lanes doesn't have a high correlation coefficient, but roads with 1, 3, 4, and 5 lanes appear to have more car accidents than with 2 lanes.



Since number of lanes and speed limit are not totally independent, I decided to look at mean accident densities for roads with a 30mph speed limit.

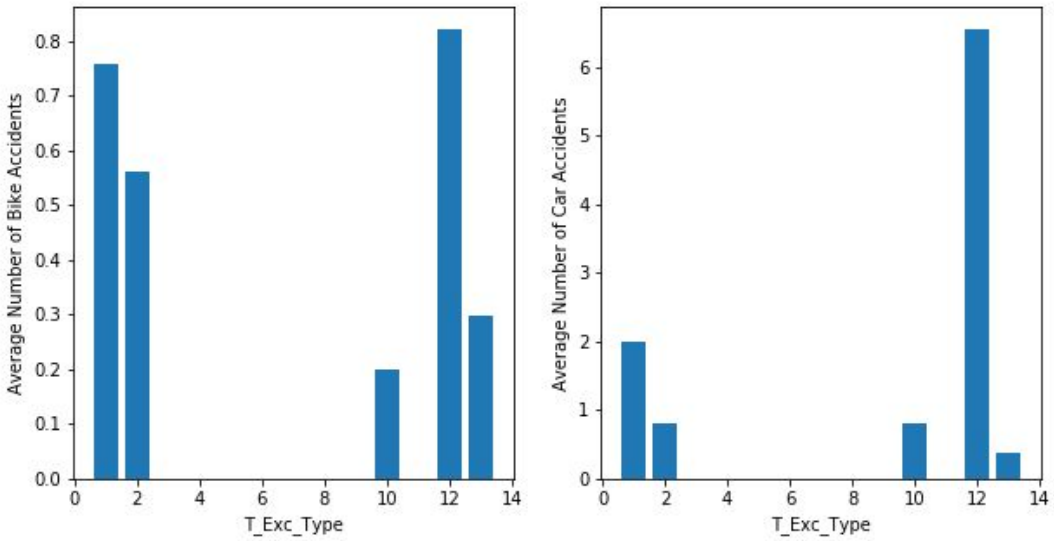
Hypothesis: roads with 1, 3, 4 and 5 lanes have a higher mean car accident density than the average car accident density

I found a p-value for rejecting the null hypothesis that there was the same car accident density for roads with 1 and 3 lanes, i.e. they had a statistically significant greater average car accident density.

Using a similar test I found that roads with 4 lanes had a higher mean bike accident density than average.

Truck Access

Higher levels of truck access does seem to have an effect on accidents, but this still may be because of what data is collected. Routes that allow hazardous trucks (Exclusion Type 12) have much higher than average number of car accidents.



The vast majority of roads have no truck exclusions, but around 5000 have some kind of exclusion.

Hypothesis: roads with any kind of truck exclusion have a lower mean accident density than roads with no truck limits

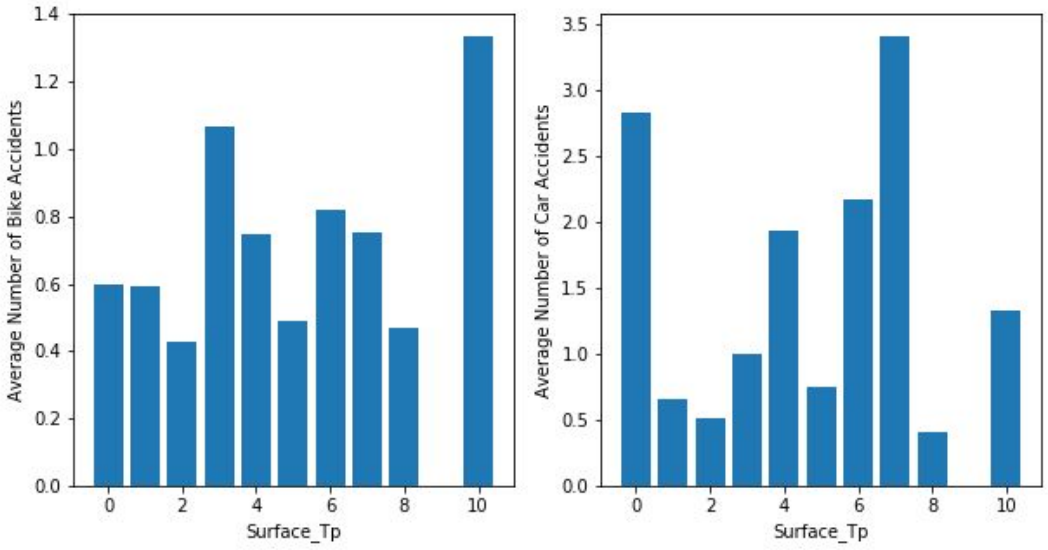
I found extremely small p-values for rejecting the null hypothesis, so truck exclusion does have a strong effect on accident density for both bikes and cars.

Hypothesis: allowing hazardous trucks leads to a higher mean accident density than average

I found a very small p-value for rejecting the null hypothesis with respect to cars, indicating that the presence of hazardous material trucks leads to a higher number of car accidents, but not enough evidence to say the same for bikes.

Surface Type

The data for surface types is not very good, but we do have some.



These graphs can be a bit misleading because only types 2, 4, 5, 6, and 7 have much data at all.

Surface Type	**Number of Data Points**
-1 = Unknown	2051
0 = Unknown	8743
1 = unimproved, graded earth or soil surface road	49
2 = gravel or stone road	960
3 = brick road	73
4 = block road	392
5 = surface-treated road	7895
6 = bituminous concrete road	53584

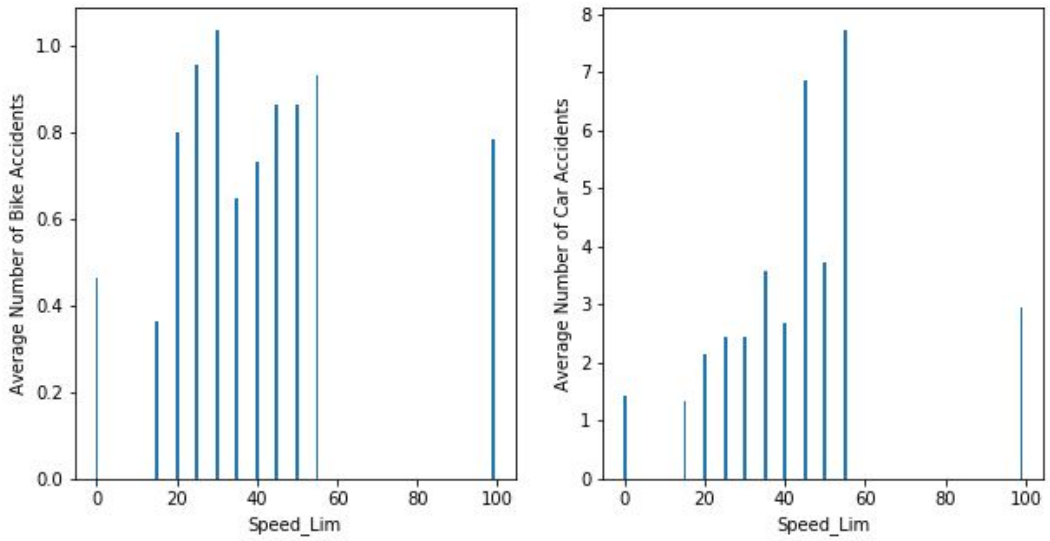
7 = portland cement concrete road	337
8 = Composite road, flexible over rigid	34
9 = Composite road, rigid over flexible	1
10 = stone dust	3

Hypothesis: Road types 2, 4, 5, and 7 have higher mean accident densities than type 6

The p-values support this hypothesis for higher bike and car accident density for types 2, 5, and 7 but not 4.

A Deeper Dive Into Speed Limit

Speed limit doesn't have a high correlation coefficient, but looking at the bar chart shows the number of bike accidents may be growing from 15 to 30, and then again from 35 to 55. Car accidents mostly increase from 20 to 55.



Hypothesis: the mean car accident density increases from 20 to 30 mph, as does mean bike accident density. (I'm starting with 20 mph because there weren't many data at 15 mph)

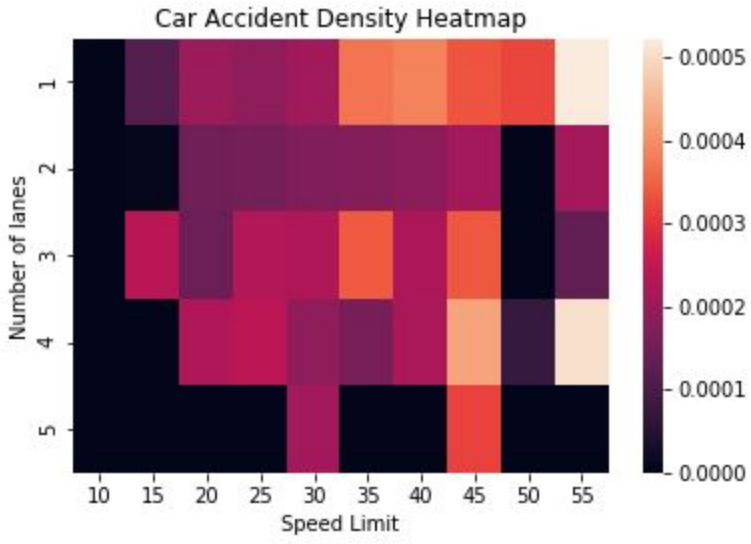
I found the p-value only showed statistical significance for bike accidents. The p-value for rejecting the null hypothesis that there is no difference between bike accidents in:

- 20 mph zones and 25 mph zones is 0.0009764585459075452
- 25 mph zones and 30 mph zones is 0.0006733426655266123
- 20 mph zones and 30 mph zones is 9.245160606207875e-11

Toward Machine Learning

I have identified some road attributes that are correlated with higher accident density for both bikes and cars.

The next step is to use machine learning to find out if the data can be used to predict accident density, especially where there might be multiple variable interactions. For instance, during exploratory data analysis I found evidence that increasing the speed limit on a one lane road had a correlation with increased car accidents.



There may be many other interactions between variables that could be interesting but could only be found with more advanced machine learning.

Machine Learning

For more detail see the full [Machine Learning report.]

Data Re-Wrangling

The Data Wrangling performed earlier was sufficient for exploratory data analysis and visualization, but for machine learning, it needed to have missing values filled in, and to have classification variables turned into dummy columns.

Finding Effective Models

I calculated number and density of bike and car accidents in each region where they occurred. I hoped that I would be able to use regression to create a model to predict number of accidents based on road attributes. I used many different regression algorithms, but never produced a sufficient R^2 score, approaching 1.

This table shows the best R^2 scores for regression.

Model	**Vehicle**	**R^2 Score**
RandomForestRegressor	car	0.109
RandomForestRegressor	bike	0.451

The best these did was accounting for 45% of observed variance, which is not enough to make a very useful regressor. I did try scaling the data to see if that would help, but it did not.

Classification of roads that are likely to have accidents versus those that are not is still a useful project. Next I created a new response variable that was 0 if no accidents occured in the road area and 1 if it did. I used many different classification algorithms and found that some produced almost 90% accuracy.

Model	**Vehicle**	**Accuracy**
DecisionTreeClassifier	car	0.886
DecisionTreeClassifier	bike	0.902
RandomForestClassifier	car	0.885
RandomForestClassifier	bike	0.908
BaggingClassifier	car	0.895

BaggingClassifier	bike	0.913
GradientBoostingClassifier	car	0.885
GradientBoostingClassifier	bike	0.895
AdaBoostClassifier	car	0.859
AdaBoostClassifier	bike	0.884

For each of these, I examined the confusion matrix as well, to make sure that the false positives and false negatives were fairly balanced, though for this problem, false positives are better than false negatives. The safety of a given road can always be improved.

For more detail see the following notebooks:

- Logistic Regression
- K Nearest Neighbors
- SVM Classifiers
- Tree and Ensemble Classifiers

Hyperparameter Tuning

Since the Decision Tree performed both fast and pretty well, and Random Forest, Bagging, and Gradient Boosting, all performed well, I decided to tune these.

Model	**Vehicle**	**Pre-Tuning Accuracy**	**Best Tuned Accuracy**
DecisionTreeClassifier	car	0.886	0.888
DecisionTreeClassifier	bike	0.902	0.905
RandomForestClassifier	car	0.885	0.897
RandomForestClassifier	bike	0.908	0.919
BaggingClassifier	car	0.895	0.895
BaggingClassifier	bike	0.913	0.922
GradientBoostingClassifier	car	0.885	0.894
GradientBoostingClassifier	bike	0.895	0.913

For more details on Hyperparameter Tuning, see this notebook.

The best performing model for predicting car accidents was a Random Forest Classifier. The confusion matrix for the best tuned model is shown here:

	Actual No Accident	**Actual Accident**
Predicted No Accident	11423	1060
Predicted Accident	1242	8512

The classification report for the car accident Random Forest Classifier:

	precision	**recall**	**f1-score**	**support**
No Accident	0.90	0.92	0.91	12483
Accident	0.89	0.87	0.88	9754
Accuracy			0.90	22237
Macro Avg	0.90	0.89	0.89	22237
Weighted avg	0.90	0.90	0.90	22237

The best model after tuning for predicting Bike Accidents was the Gradient Boosting Classifier. Here is the confusion matrix:

	Actual No Accident	**Actual Accident**
Predicted No Accident	17173	353
Predicted Accident	1586	3125

And here is the classification report. This has a low recall for accidents, and a high false-positive rate for predicting an accident. However, given that roads can always be made safer and an area with no accidents may be an area that simply hasn't had an accident yet, I think this may be a benefit to the model.

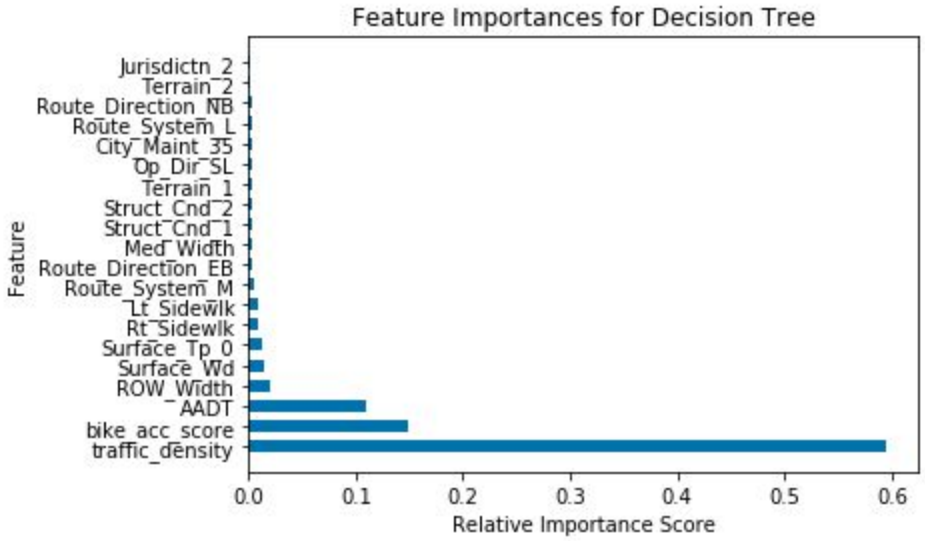
	precision	**recall**	**f1-score**	**support**
No Accident	0.92	0.98	0.95	17526

Accident	0.90	0.66	0.76	4711
Accuracy			0.91	22237
Macro Avg	0.91	0.82	0.85	22237
Weighted avg	0.91	0.91	0.91	22237

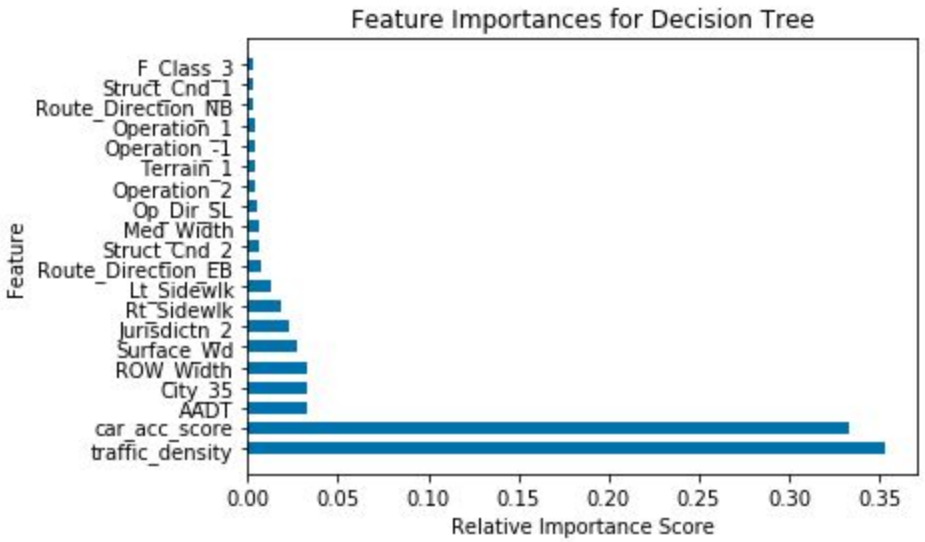
Feature Selection

I used the decision tree feature importance to identify the most important features for predicting car and bike accidents.

Most important features for predicting car accidents:



Most important features for predicting bike accidents:



Some of the features are dummy columns for categorical features, so I kept not just the top categories, but all the dummy features. It is still a much reduced set. For cars, it went from 210 features to 77. For bike accidents it went from 210 to 54.

Below is a table comparing the accuracy of running these models on a reduced feature set. These are untuned accuracies.

Model	**Vehicle**	**Pre-Tuning Accuracy**	**Best Tuned Accuracy**	**Reduced Feature Accuracy (untuned)**
DecisionTreeClassifier	car	0.886	0.888	0.888
DecisionTreeClassifier	bike	0.902	0.905	0.810
RandomForestClassifier	car	0.885	0.897	0.892
RandomForestClassifier	bike	0.908	0.919	0.813
BaggingClassifier	car	0.895	0.895	0.893
BaggingClassifier	bike	0.913	0.922	0.812
GradientBoostingClassifier	car	0.885	0.894	0.883
GradientBoostingClassifier	bike	0.895	0.913	0.801

We can get almost the same accuracy with fewer features for predicting car accidents, but decreasing features reduces the accuracy for predicting bike accidents by quite a bit, from a best of 92% to 80%.

Exploring the Models

In this section I will dig into the implications of the machine learning models. For code and more detail, see this notebook.

Model Sensitivity

Machine learning models can be used to predict how changes in features might predict changes in outcome. Since traffic is something that can change easily, I decided to look at the effect of increased traffic on predicted bike and car accidents.

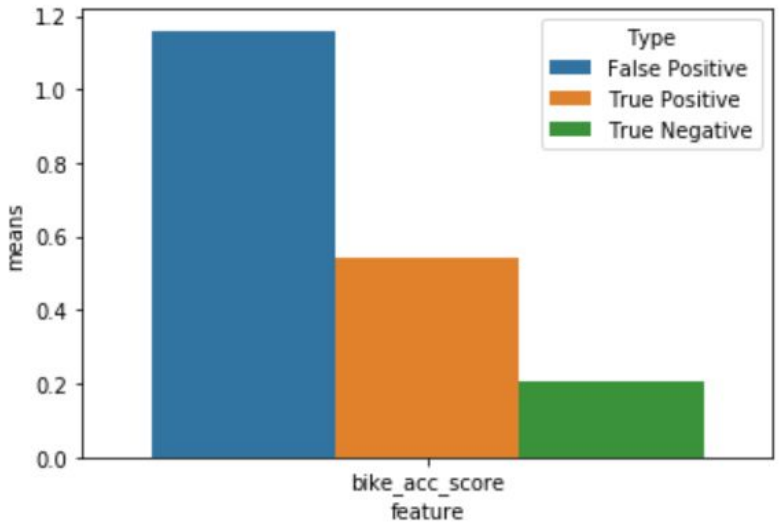
Unfortunately, I found that small changes to traffic led to huge increases in predicted accidents, for instance, 1% more traffic led to 82% more car accidents predicted and 151% more bike accidents.

However, because many of the features aren't actually independent, changing one thing like traffic isn't very accurate to what would happen in reality. An area with more traffic, i.e. more cars passing, probably has a higher speed limit, and larger road surface, so changing one variable without changing the others isn't that illuminating.

Exploring the Confusion Matrix

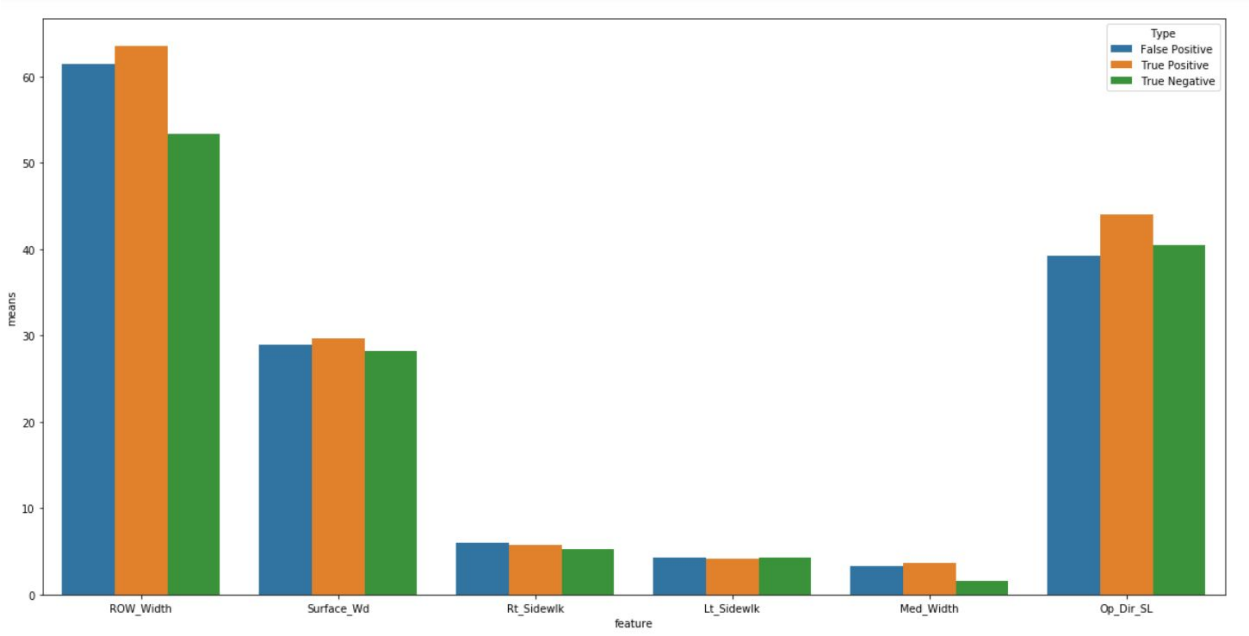
The confusion matrices indicate how often the model predicts true and false negatives and positives. Because false positives may be places where accidents simply have not occured yet, I wanted to explore similarities between true and false negatives.

I constructed a data set that compares a few features for true positives, false positives, and true negatives. One finding from this exploration was that false positive car accidents tended to have high numbers of bike accidents:

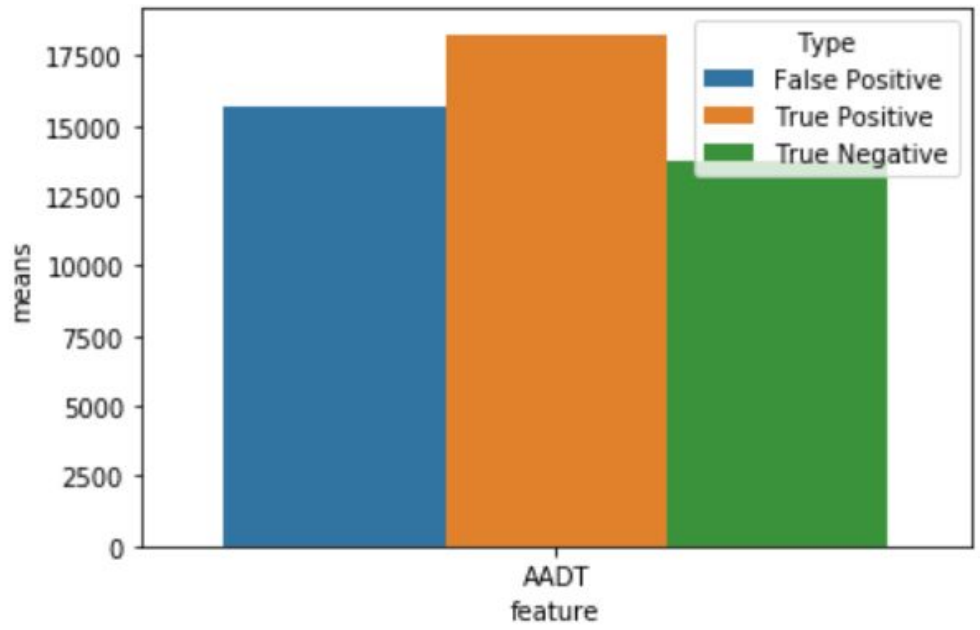


This may indicate locations that could have a car accident in the future.

Graphing some other quantitative features shows that higher right of way width and median width may indicate places that will have a car accident in the future:



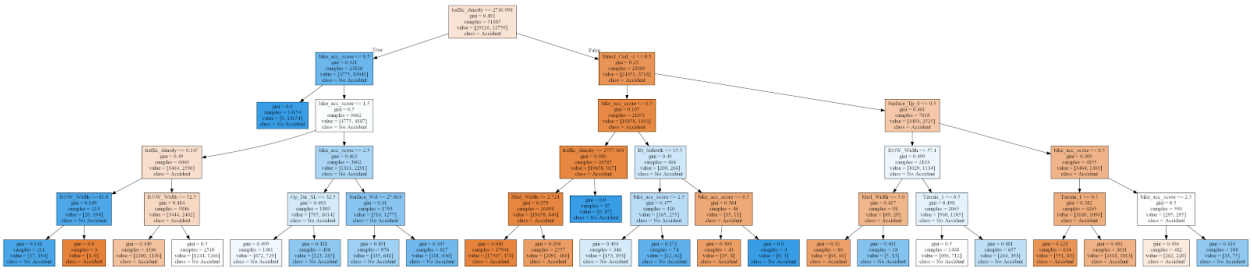
Performing a similar construction on shows false positives have higher traffic than true negatives, indicating places which may have a bike accident in the future:



Visualizing the Decision Trees

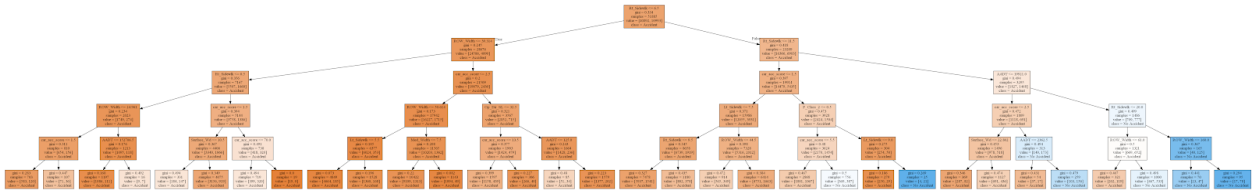
Single decision trees were pretty effective for predicting whether an area would have accidents or not, so I wanted to visualize these trees. They could be used by a road planner to make decisions that would affect the likelihood of accidents.

Here is a decision tree for predicting car accidents that only has 5 layers and 12 features considered, and still gives 87% prediction accuracy.



This decision tree shows how different choices lead to a higher likelihood of car accidents. A road planner could use it to make decisions about how to construct a road, or create signage that might decrease accidents.

Similarly, here is the decision tree for predicting bike accidents based on a reduced feature set. This can only give a 79% prediction accuracy, but it still might be helpful:



Conclusions and Potential Next Steps

The feature selection graphs and trees show that the second most important feature for predicting bike accidents is car accidents, and vice versa. This supports my initial thought that making an area safer for bikes might make it safer for cars.

Traffic volume is the most important feature for predicting both types of accidents. This argues for increasing public transportation to decrease traffic.

Some potential next steps include:

- Collecting more recent bike accident data, since this data set was collected in 2012
- Seeing if the model predicts car accidents for areas outside Boston, or if Boston road conditions are unique
- Creating models for other cities and seeing if they support similar conclusions