Data Science Machine Learning Take Home Assignment

Anton Sitkovets
asitkovets@gmail.com

Introduction

The E-Bike Survey Response Results data set provides us with a CSV formatted survey giving information on Torontonions and their knowledge on rules about E-Bike laws within Toronto. The purpose of this exercise is to see if we can create a model that can predict whether an individual would answer “No - I do not have access to a private motorized vehicle” to the question “Does your household have access to any of the following types of private motorized vehicles?”. The result of this question has a multinomial response, but since we only care about one response we can treat it as a binomial class label. We can use a supervised learning model to decide whether someone belongs to class label X or NOT X.

Which models did you consider? Which Model did you choose and why? How good was it?

The three models that I considered for the task were Naïve Bayes, Logistic Regression, and Support vector machine(SVM).

The Naïve Bayes Classifier is a generative method that uses the assumption that the features are conditionally independent given the class label. Given this assumption, Naïve Bayes can avoid issues with scale as the model does not scale exponentially with the number of features. Although we assume the features to be independent, this is usually not the case for all pairs of features. But despite this fault in the assumption, Naïve Bayes classifiers are easy to implement and tend to work well because the model is simple with only O(CD) parameters for C classes and D features, leading to little overfitting. Since the features have a multinomial distribution, a Dirichlet distribution can be used to represent the likelihood of the feature being labeled as class c. We can then check the posterior value for log p(y=”owns motor vehicle” | x) and if the result is greater than log ½, the input likely belongs to label “owns motor vehicle” and if less than log ½ it belongs to “does not own motor vehicle”. I am most familiar with this classification model as I worked on a similar problem in my machine learning class about classifying mushrooms for being poisonous or edible.

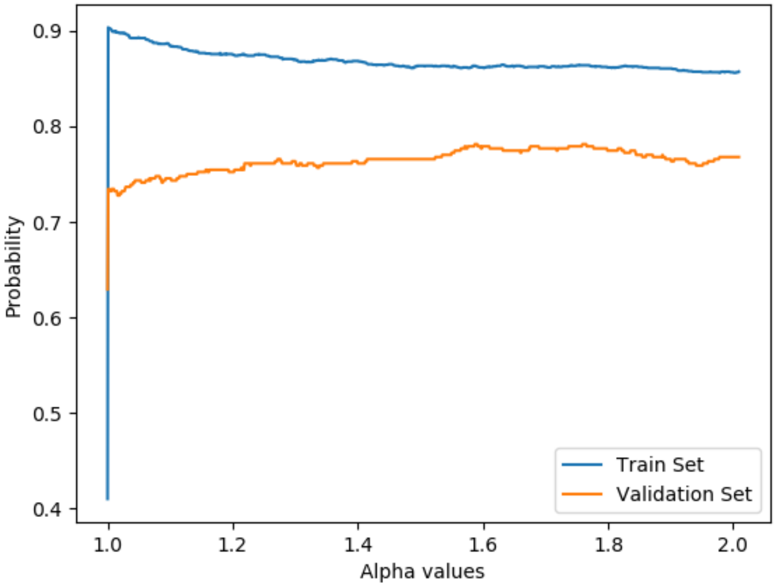
The Logistic Regression model is a discriminative method which applies a sigmoid function to a linear decision boundary. We can formulate this decision boundary using the equation $a(x) = b + w^T x$, where b is a bias and w is a weight vector. To find the weight vector w and bias b, we must run the gradient descent algorithm to find the parameters that minimize the log likelihood of the data. The benefits of logistic regression are that it is quick to train, fast at classification, has good accuracy for simple data sets and resistant to overfitting. The one con to logistic regression is that linear decision boundaries are too simple for more complex problems.

The Support Vector Machine (SVM) model is a discriminative classifier which given a set of training data, generates an optimal hyperplane for categorizing new examples. The hyperplane can be considered as a decision boundary between the two class label outputs. The idea is to find a plane that is as far as possible from all the points to reduce noise sensitivity, this operation is called maximizing the margin. The hyperplane can be represented using the equation $w^T x + b$. By plugging in the value of x, if the result of the equation is >= 1, then it is predicted as class 1 and if the result if <= -1, then it is predicted as class -1. If the result leads to $-1 < y < 1$, then this is of an undefined class. To find the values of w and b, the solution must correctly classify the training examples and maximize the margin. The benefits of SVMs are that they are guaranteed to have the optimal solution and have good accuracy, while the cons include that training time on large datasets can be high and it is a less effective method on nosier data with overlapping classes.

I ended up choosing the Naïve Bayes approach as it is the model that I had the most experience working with. I could have just done the experiment with any of the models using a library like scikit-learn, but I wanted to show that I understand how the algorithms work. Since I was most familiar with Naïve Bayes from my experience implementing the algorithm at school, I chose this option over the other more accurate options. Per this paper (https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf), all the other models I mentioned have better accuracy than Naïve Bayes. But seeing as I am less familiar with the other algorithms and I wanted to show my programming abilities, I decided on Naïve Bayes.

Looking at the code, I first prepared the data by taking the csv table and converting it to a matrix of integer values. I also kept an Ordered Dictionary mapping the feature value to the integer associated to it. I then split the data up 80/20 for training/test data. Next, I created two separate matrices for data with people who don't own a motor vehicle and for people who do. Using these matrices, I found the prior probabilities of the two outcomes. Next, I plotted a comparison of the distribution of people who don't own motor vehicles and those who do for each feature. I then calculated the likelihood of each feature value to the class label using the MAP estimate for the Dirichlet distribution. Finally, I used the log sum exp trick to find the posterior value of the prediction and checked the result of the prediction with the actual label to calculate the accuracy of the model. I did this over a range of values for the Dirichlet hyper parameter values. Lastly, I plotted the results of the prediction accuracy for all the hyper parameter values for predicting the training data and the test data.

The Naïve Bayes approach resulted in a best of 90.34% accuracy for classifying the training set and 78.12% accuracy for classifying the test set. Using one of the other classification techniques listed above could have improved the accuracy of this problem, but Naïve Bayes proves to be a fast and effective classification technique.



Best accuracy for training set was 90.3405918481%
The best alpha value for training data is 1.001
Best accuracy for test set was 78.125%
The best alpha value for test data is 1.587

What was the pattern of missing values? Was it random? Could those be inferred from the context?

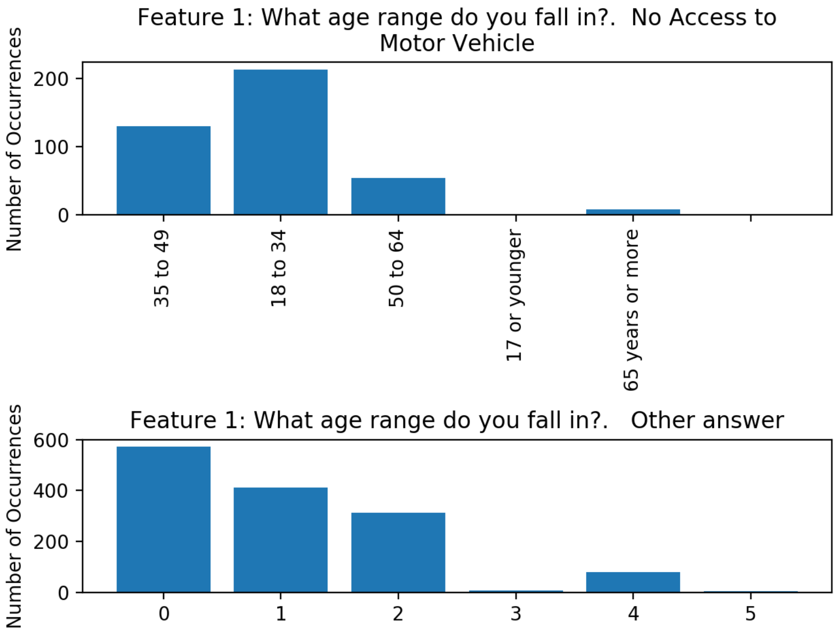
Looking at the data set, the only pattern I can see about the missing values is that people would prefer not to answer more personal questions about themselves in government surveys. The only columns that I can see with missing values are age, sex, health, education, household income, and employment status. From this I can infer that the survey left these questions as optional to give individuals taking the survey a sense of privacy, whereas the rest of the questions are mandatory and more about commuting and e-bike laws.

The distribution of missing values among these optional questions seems random in the context of the entire data set, but there must exist some inferred relationship to why someone with a post graduate degree and who is self-employed would leave out their salary. One technique we can use to work with the data would be to simply discard data with missing values, but this would be a waste of useful information for all the other columns that have data and may discard a high portion of the data. Another technique would be to ignore features that have missing values entirely. This would mean that we wouldn’t even consider the optional questions in our model. Although this can be done, we would be missing out on some significant features such as income and age.

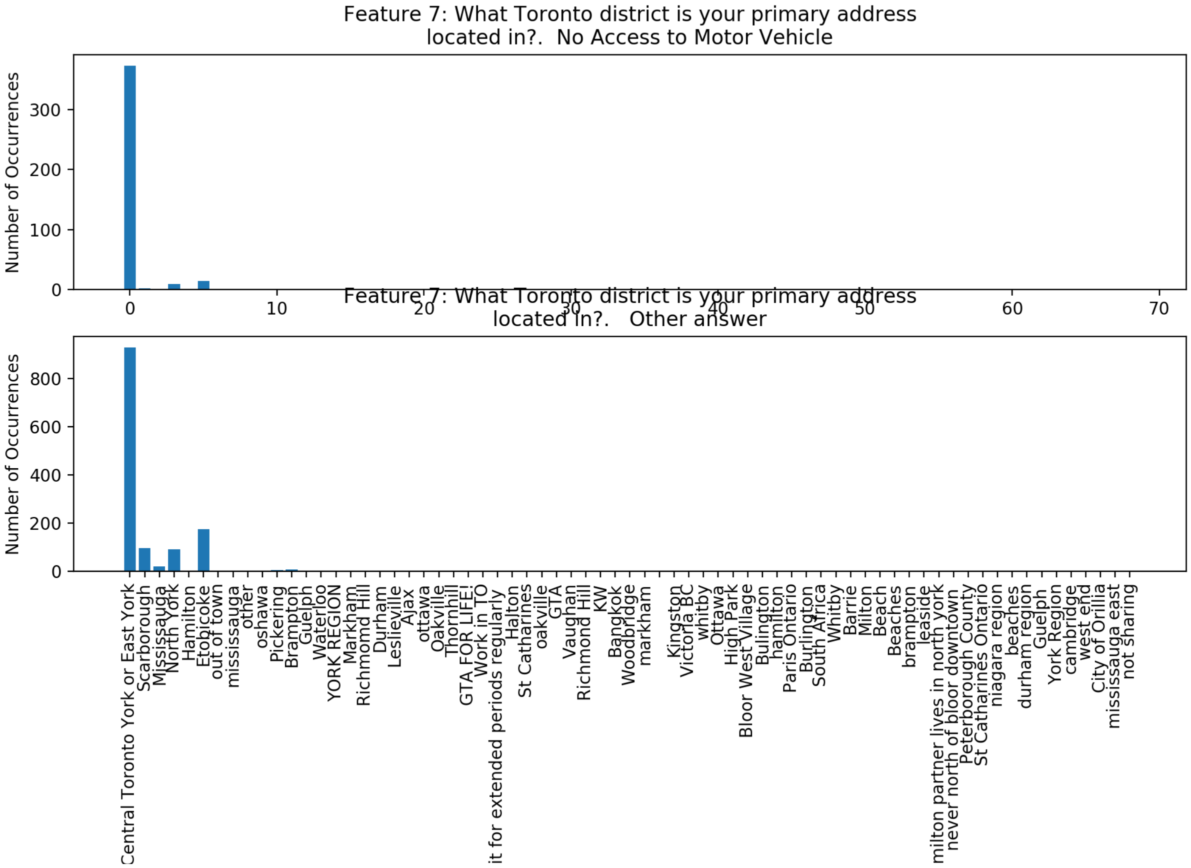
We can attempt to find these missing values by several means. We could simply assign the missing value as the value that occurs most frequently in the given set of data(mode). This is a crude solution that does not consider all the other information that we know about the individual. Another possible method would be to train a Naïve Bayes classifier for each of these optional questions and then we can classify the data with missing values per their answers to the other questions. This would be fairly accurate but would take a very long time as we would need to train a different model for each of the 6 additional questions and then run the classification on the data with missing values. The key here is to find features that are correlated and try to predict the most likely value for the missing value based on other pieces of data with the same answer to the correlated data. For instance, we can say that education and income are correlated, and if someone chooses not to input their income but has given their education, we can ball park their income based on other data. Upon further research, I found that if we try to infer random missing data, it could incorrectly bias the results.

Which features were significant in predicting the target response?

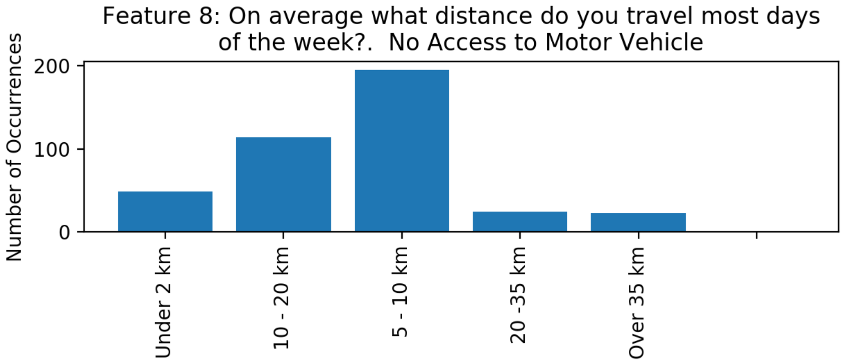
Looking at the generated histograms, the features that were significant in predicting the target response were: age range, address location, distance traveled per week, income, and transportation option most used.

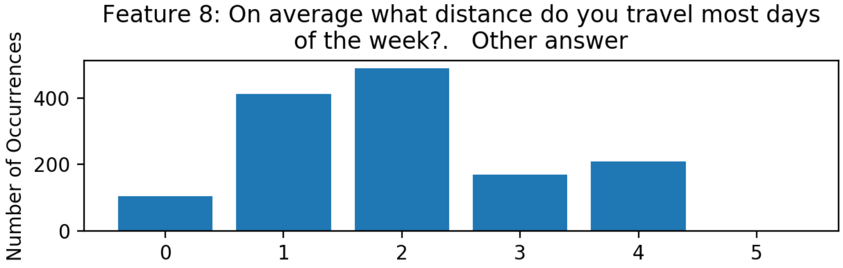


From the age range feature, we can say that in the “no motor vehicle” distribution, the likeliest age range to not own a vehicle 18 to 34. Whereas the likeliest age range to own a car is 35 to 49.

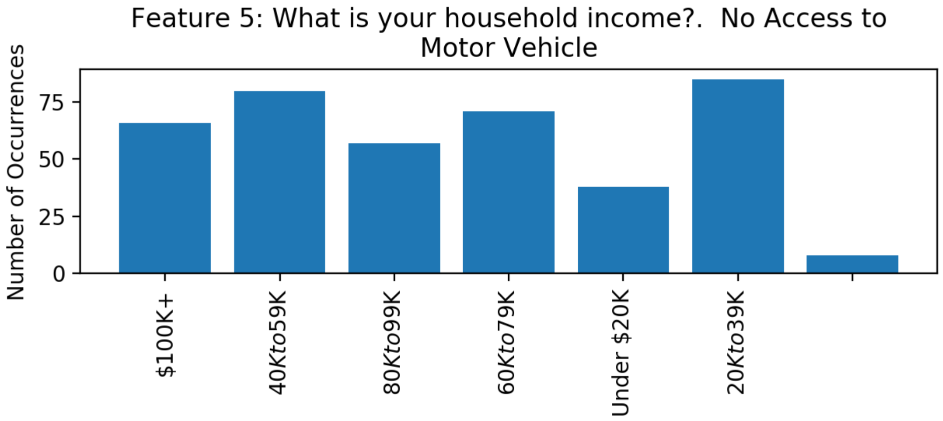


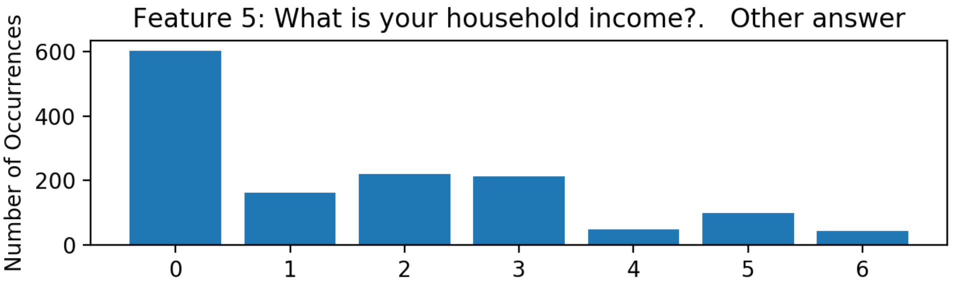
As expected, individuals that live further away from the downtown core are more likely to own motorized vehicles. As can be seen from the lower graph as people from Scarborough, Mississauga, North York and Etobicoke are more likely to own motor vehicles than not. For individuals that live in Central Toronto or East York, they are still more likely to own motor vehicles, but among those that do not own motor vehicles, most live in this area. From this feature, we can say that anyone living outside of the central Toronto area is more likely to own a motor vehicle.



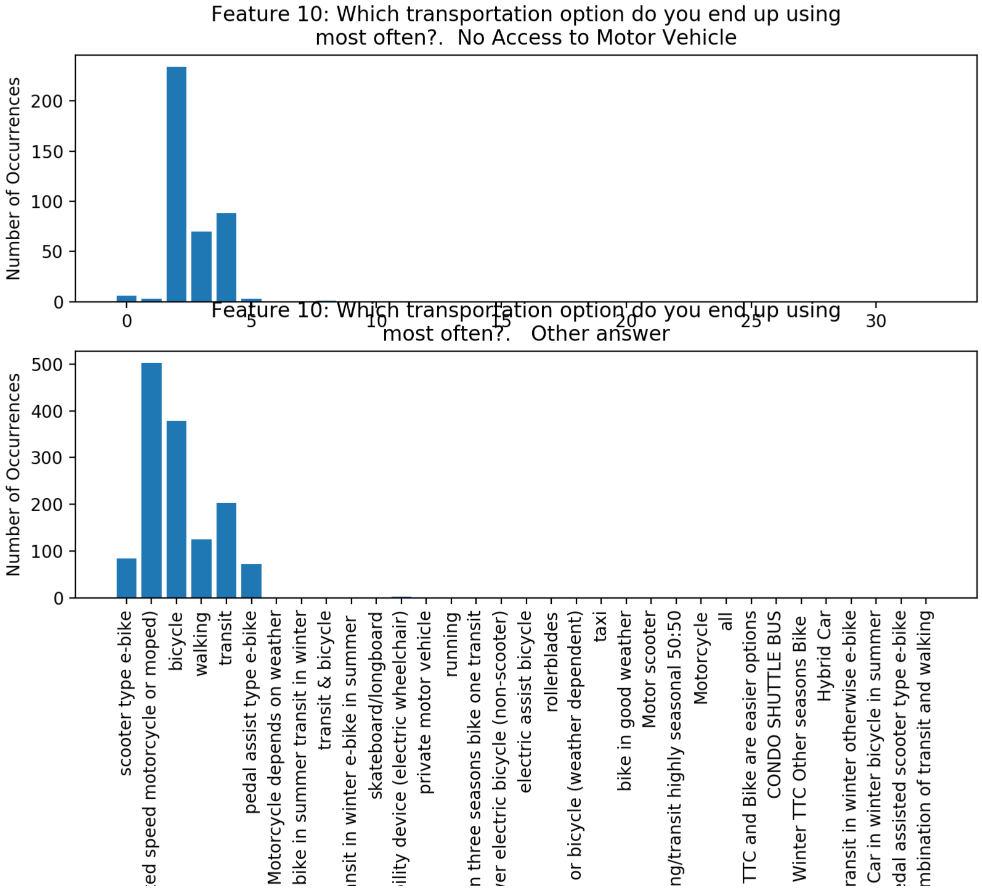


The distributions for car owners and non-owners are similar except for when someone travels 20 or more km on a given day. As expected if someone needs to travel this extensively each day, they are more likely to own a motor vehicle. Whereas the distribution for traveling 10 km and under each day is similar for both histograms.





Income plays a large role in the prediction for motor vehicle ownership as it makes sense that someone should not own a car if they cannot afford it. Surprisingly there is a similar number of people who don't own a car for each income bracket. Households with large income are much more likely to own cars and households with low income are more likely to not own cars.



Transportation option used most is a useful feature for prediction as it is clear that someone without a motor vehicle wouldn't choose the "Car, motorcycle, etc." option and would mostly choose biking, walking or transit. Although surprisingly many car owners end up using bicycles for the most part.

Surprisingly, features such as length of commute, employment and health did not prove to be significant features for training the model. One would expect an individual's health level to be relevant to how they travel. But it looks like people who don't own motor vehicles are as healthy as those who do or just say that they are.

If you could re-design the survey for next year, what question(s) would you add or remove in order to improve the precision of the prediction?

If the goal of the survey was to create a system that can predict if someone owns a motor vehicle, I would remove all the open-ended questions that allow for user responses. These questions lead to features having too many possible values and do not generate significant features for classification.

I would also ask, “What kind of outdoor activities are you most involved in?”. This question would be very discriminative on whether someone owns a car if they say they like to ski, fish, camp or hike as these all require a car to take part in.

Another new question would be, “How far is your work from your house?”, as people who work far from their home are more likely to own cars. Although people who work close to their home can also own cars, the case for when someone works very far from their home is a strong distinguishing result.

Can also ask “What industry do you work in?”. Some industries such as construction, repair service, technician, etc. require you to own a car to reach different work sites.

“Are you satisfied with your local public transit system?”. In my opinion the main reason anyone would get a car is because the public transit in their city is unreliable and slow. If someone finds their public transit unsatisfactory, then they are more likely to own a car.

But since these questions are completely unrelated to the main point of the survey of finding out peoples’ opinion on ebike laws, I don’t think these questions will help anyone except us. The only relevant question for the original purpose of the survey would be “How far is your work from the house?”, although this is like asking “How far someone travels in a day in a given week”, it is more specific to allude to whether the person requires a car or not and gives a reason as to why they would need one.