Shopify Data Science Machine Learning Task Home Assignment

**Problem:**

Based on [E-Bike Survey Response Results](https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#88c2d085-3aba-eeef-3fb9-d7299912ad37) also available in [CSV Format](https://github.com/samtalasila/e-bike-survey-response-results/blob/master/E-Bike_Survey_Responses.csv) data available from TO Open Data, train a model to predict whether the survey responder, answers *“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?”*.

**Question #1**: Which models did you consider? Which Model did you choose and why? How good was it?

I considered a variety of models including K-Nearest Neighbours, Random Forest, Radial Basis Function Support Vector Machine, One-vs-All Logistic Regression, a simple feedforward neural network, and a two layer deep neural network. I ultimately chose a selected ensemble of models composed of One-vs-All Logistic Regression, and the two networks for my final classification as it provided the best accuracy when predicting if a household does not have access to a private motorized vehicle. The model performed with 63.8% accuracy on the classification on a whole and 53.7% (22/41 correct prediction) on the specific classification of not owning a vehicle. I provide an in depth explanation as to why machine learning models were unable to perform better on this given data set within the exploratory analysis.

**Question #2**: What was the pattern of missing values? Was it random? Could those be inferred from the context?

There were 135 NaN values in the data set, unevenly distributed among the categories (Figure 1). The primary feature with missing information is 'What is your household income?' with 60 missing entries. I considered numerous approaches to infer with these empty values. One approach would be to replace them with a new category entitled empty data. As simple as that is, it can actually be a useful approach when there is a pattern to the questions participants left blank in relation to the classification of interest. An alternative approach would be to bootstrap these values (ie. interpolate from the data in some way to fill these values in). One method to do this is to randomly sample from the feature and replace the values. A second for real valued data is to replace all empty values with the mean. A third is to determine if there is a correlated feature that may help us predict what these values could be. What I decided to do is interpolate the empty values in relation to the average income considering the category 'What level of education have you reach?”. Details can be found within the exploratory analysis however yes they could be inferred from this column and the filled in responses were as follows: High School ($61k), College ($69k), University ($75k), Post Graduate ($80k).

**Question #3**: Which features were significant in predicting the target response?

Short Answer:

The most significant features in predicting the target response were (Figure 2):

* "Which transportation option do you end up using most often?"
* "When you use Toronto's bicycle lanes do you mostly"
* "When you use Toronto's Multi-Use Trails do you mostly"
* "How would you describe your level of physical health?"
* "What level of education have you reached?"
* "Have you witnessed a collision or conflict on a trail between"
* "What is your household income?"

Long Answer

In order to determine feature significance there are two commonly used algorithms. The first is the Recursive Feature Elimination which works by recursively removing features from a given model and building on features that remain. The results of this model return a ranking of the most significant features for the given model. I demonstrate this approach in the exploratory analysis, however during my analysis I convert the categorical variables to one-hot features, which does usefully translate to a ranked representation. Instead, I consider a second commonly used approach known as Extra Trees Classifier which summarizes the identifies the most significant features by measuring the performance of many decision trees created by randomly selecting features. The method returns a percentage output representing significance across all features of which I sum considering the number of one-hot values for each category to determine the features most influential in the classification problem (Figure 2).

**Question #4**: 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 I were to redesign this challenge for next year I would try to reduce categorical feature data wherever I could. For example: age, income, distance travel most days of the week, and commute time should all be real value answers. I would then try to ask some of the seemingly categorical questions in ways to yield real valued responses. For example, 'How would you describe your physical health?' could be phrased as, 'How many hours of physical activities do you do a week?'. I would also try to summarize the categorical data more succinctly to provide at most 5 categories per answer. This is for questions such as: 'Do you support any of the following statements' and 'When you use Toronto's Multi-Use Trails do you mostly' which have 100s of unique answers. In addition, for next year if the classification question remains the same, one could design survey questions targeted specifically towards determining if someone owns a vehicle or not by asking questions such as, 'How much do you pay approximately in gas each week?'. Also, based on our determination of the most significant features, we know which questions to keep in the survey, and which to reconsider. Lastly, a large problem of this data is there is an uneven distribution of output classifications relevant to our question of interest (Figure 3). Especially for categorical data, this leads to models overfitting. Upon reviewing the data, one reason for this is the demographic sampled for this survey is overall a poor representation of a diverse group of individuals who may or may not own a vehicle. I base this on the disproportional number of high income individuals and low number of students (Figure 4 & 5). If I were to perform this survey again I would emphasize that the survey be conducted in a variety of demographic areas to ensure a more equal distribution of output classifications.

Figure 1 – Number of NaN Entries Considering Each Feature

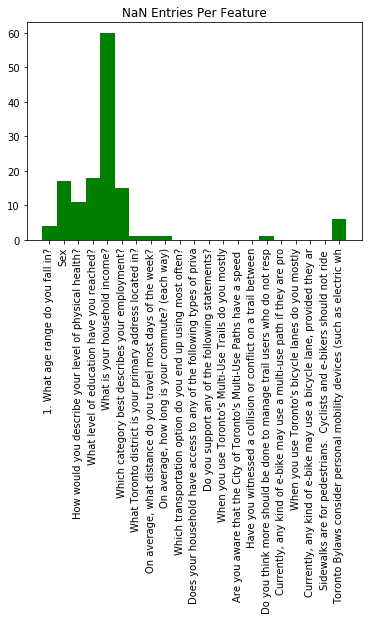


Figure 2 – Percentage Contribution of Each Feature to Correct Vehicle Prediction Classification

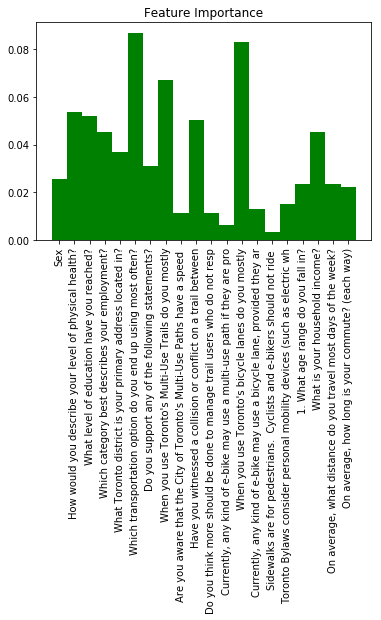


Figure 3 – Output classification requested depicting the large class imbalance.

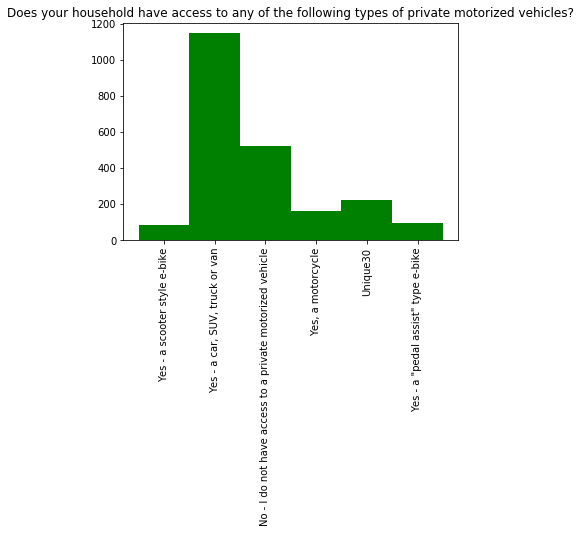


Figure 4 – Household Income showing a disproportionate number of high income families. This biases the data to include few examples of individuals who do not own vehicles.

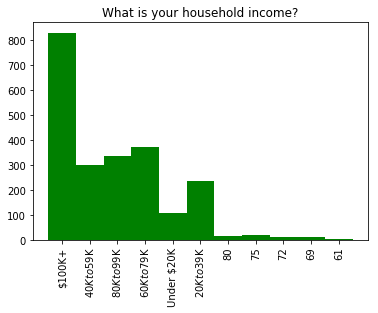


Figure 5 – Employment showing a disproportionate number of students. This biases the data to include few examples of individuals who do not own vehicles.

