INFO251 - Applied Machine Learning
Shrestha Mohanty, Udit Nakhat, Vikram Reddy, Pavan Yedavalli

Project Proposal

Predicting vehicle types and transit modes for more effective urban mobility policy and corporate advertising

Overview

This proposed project will seek to understand more closely urban and suburban mobility patterns in the San Francisco Bay Area, specifically utilizing machine learning to understand what types of automotive vehicles certain people tend to use. This project aims to integrate urban data science with behavioral choice modeling in order to inform future government policy and corporate advertising. By understanding vehicle choice based on personal

attributes, policy and advertising can be more directed. For instance, if it is known that males between the ages of 40 and 50 with six-figure incomes tend to drive Teslas over all other vehicles, this information could be useful to Tesla for targeted advertising. In a second example, if it is known that most hybrid vehicles are used by women from Berkeley who tend to use their cars only for leisure, then perhaps the City of Berkeley could incentivize policy in such a way that would encourage greater hybrid adoption and parking incentives. This project aims to address both the prediction of vehicle type as well as interpreting which features most impact vehicle type.

The extensive, publicly available California Household Travel Survey (CHTS) dataset (collected from 2016 to 2017), utilized heavily by the Urban Analytics Lab (UAL) at Berkeley, will provide the raw data from which the model will be created. Previous work in the space has included understanding location preferences in housing and land use, and using spatial characteristics of the data to visually understand trip patterns. The researchers in the UAL have focused on activity-based modeling, specifically understanding behavior in dense urban corridors, but no work, to the best of our knowledge, has addressed preferences of travel modes and vehicle types, which can have significant implications in policy and advertising. This novel research is relevant to urban mobility emphases in the Transportation Research Board (TRB), Association of Computing Machinery (ACM), and/or Institute of Electrical and Electronics Engineers (IEEE) journals.

Methods

Preprocessing and Descriptive Statistics:

The first step of the project is to understand and organize what data have been collected in the CHTS dataset. The data are organized based on `households, places, persons, activities, vehicles, and trips`. This project is associated with all of the data except the `trips` data, which involves a scope that is too great for this analysis.

In the exploratory data analysis, we begin by individual assessment and preprocessing of each dataset. Within the `households` dataset, we first remove all columns that are neither relevant nor have comprehensive data (e.g. contain 'redacted' data). We can see from the correlation matrix that there are several features that seem to have correlation, such as the `hispanic_flag` and `interview_language` features, which tends to make sense given that most households that have the `hispanic_flag` marker may be more inclined to do an interview in their native language. This may allow us to eliminate one of these columns to reduce computation expense and avoid errors in collinearity.

The `vehicles` dataset contains the actual target variable, `veh_type`. We also see certain trends in this dataset itself, such as the make of the vehicle and the fuel type, which makes sense, for instance, when certain trucks require diesel. In Figure 1, we show the joined version of the two datasets.

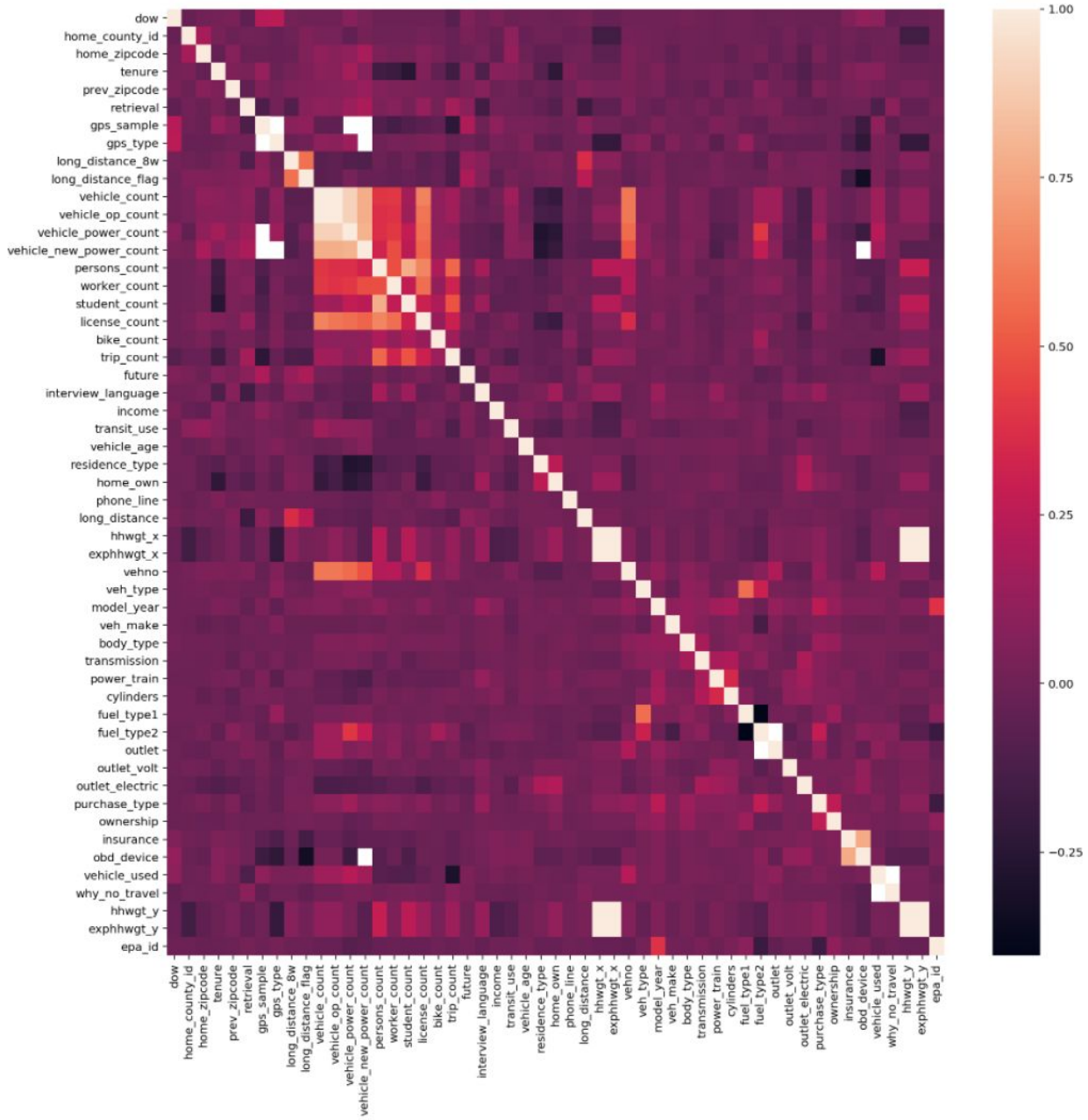


Figure 1. Correlation Matrix for households and vehicles joined dataset

The `persons` dataset is perhaps the most pertinent one from which we will glean the most information regarding specific attributes of a person. As shown in Figure 2, in its own

correlation matrix, after preprocessing, we can see there are many interesting correlations already, such as employment industry and whether the person takes a toll road. After fitting the appropriate ML models, we can perhaps use this insight for later.

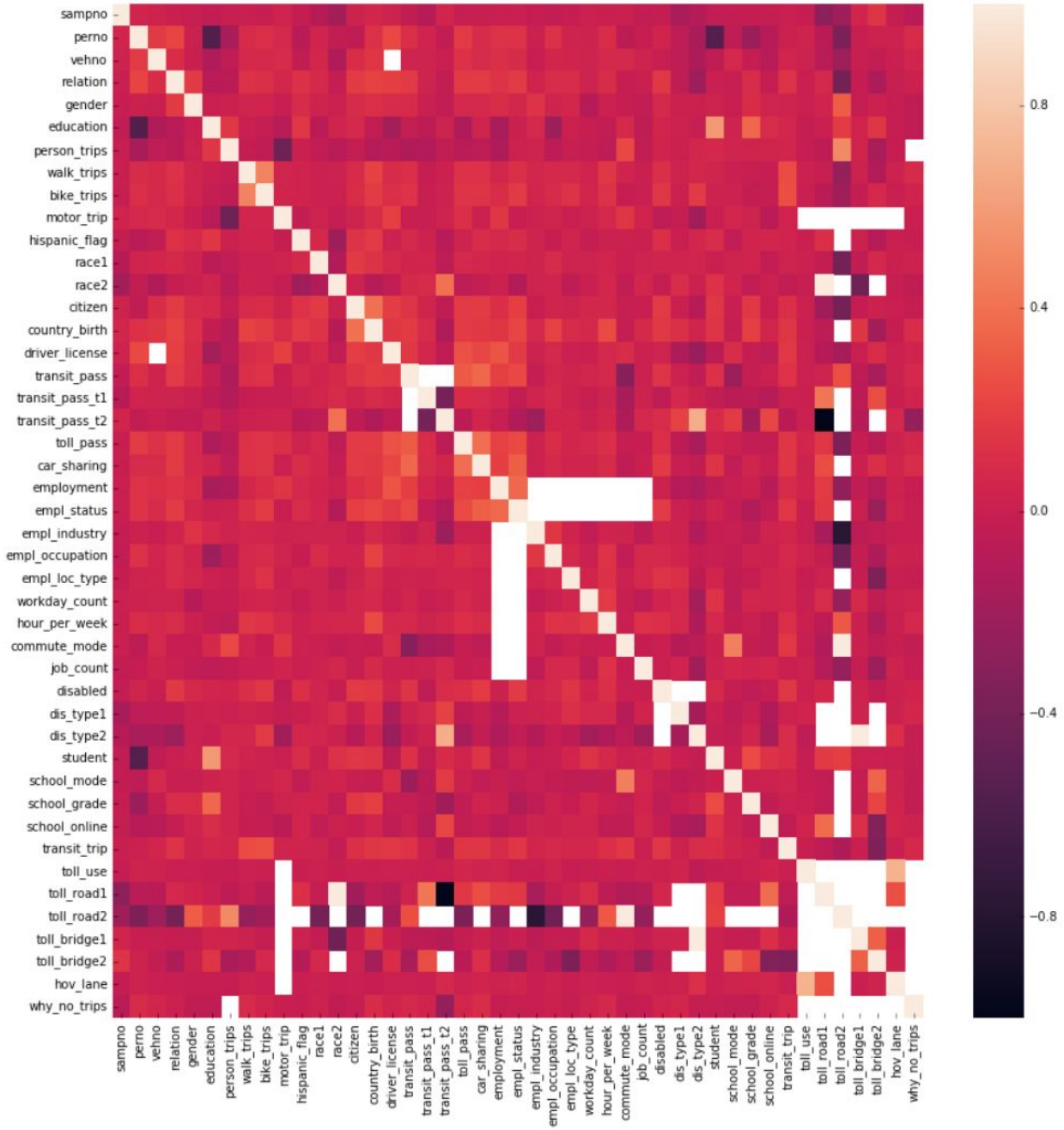


Figure 2. Correlation Matrix for Persons dataset

These are just a few examples of the datasets that we joined together in a larger resulting dataframe, after cleaning and preprocessing the data. We also developed functionality to create subset dataframes for easier tests and sanity-checks.

Dimensionality of the Data:

The California Household Travel Survey dataset has 56 datasets, out of which we are using the following three datasets that we felt were relevant to our project and would be helpful in predicting the vehicle type. The columns serve as features to our machine learning models. We have provided dimensionality along with basic statistics summary of a subset of pertinent variables for each of the datasets.

Survey Persons:

Dimension: 109,113 rows x 152 columns

Description:

This dataset contains demographics information such as education level, age, gender, and income of the respondents of the travel diary survey.

Basic statistics:

	max	mean	min	std
age	REDACTED	NaN	REDACTED	NaN
bike_trips	99	3.024788	0	13.605034
car_sharing	9	2.009035	1	0.449909
citizen	9	1.202989	1	0.585213
commute_mode	99	6.924342	1	5.820453
country_birth	9999	2774.229022	1919	2385.320605
dis_type1	99	30.067809	1	42.236533
dis_type2	98	25.552189	1	39.068523
disabled	9	1.962991	1	0.445913
driver_license	9	1.107163	1	0.442123
education	9	3.744723	1	1.987373
empl_city	NaN	NaN	NaN	NaN
empl_industry	99	57.655808	11	18.317656
empl_loc_type	99	1.343713	1	5.005868
empl_occupation	99	30.052695	11	20.680663
empl_primarycity	NaN	NaN	NaN	NaN
empl_status	99	3.691944	1	10.022949
empl_zipcode	NaN	NaN	NaN	NaN
employment	9	1.390152	1	0.615789
gender	9	1.539944	1	0.673481
hispanic_flag	9	1.926439	1	0.819034
hour_per_week	999	73.470499	1	183.933954
hov_lane	99	3.739879	1	13.041183
job_count	9	1.370140	1	0.724290
motor_trip	2	1.266158	1	0.441957
perno	8	2.061779	1	1.176517

Figure 3. Basic statistics - persons dataset

Survey Households:

Dimension: 42,426 rows x 83 columns

Description:

This dataset includes data from the households that participated in the travel diary survey. It contains information such as the location and the number of persons in the household. We also use this dataset to filter the houses in the San Francisco Bay Area using the home county id.

Basic statistics:

	max	mean	min	std
assn	13132	1.274002e+04	12532	144.468665
bike_count	99	1.761709e+00	0	3.697811
dow	7	3.620793e+00	1	1.705850
future	2	1.309727e+00	1	0.462405
gps_sample	3	1.712198e+00	1	0.794858
gps_type	3	1.149723e+00	1	0.508323
hispanic_flag	2	1.840968e+00	1	0.365725
home_block_id	REDACTED	NaN	REDACTED	NaN
home_city	YOUNTVILLE	NaN	ALAMEDA	NaN
home_county_id	97	5.710026e+01	1	36.743666
home_lat	REDACTED	NaN	REDACTED	NaN
home_lon	REDACTED	NaN	REDACTED	NaN
home_own	9	1.226557e+00	1	0.513193
home_own_other	NaN	NaN	NaN	NaN
home_primarycity	YOUNTVILLE	NaN	ALAMEDA	NaN
home_state	CA	NaN	CA	NaN
home_tract_id	6.09715e+09	6.057452e+09	6.0014e+09	36733322.374210
home_zipcode	96141	9.462943e+04	93308	434.021297
income	99	1.511745e+01	1	27.796747
interview_language	2	1.029130e+00	1	0.168180
license_count	7	1.858878e+00	0	0.799135
long_distance	9	1.525373e+00	1	0.870590
long_distance_8w	9	1.542460e+00	1	0.565936
long_distance_flag	2	1.190677e+00	1	0.392887
noveh1	99	5.618421e+00	1	9.633459
noveh2	12	5.193370e+00	1	3.243996
persons_count	8	2.472568e+00	1	1.267717
phone_line	99	1.486670e+00	0	5.707645
prev_city	NaN	NaN	NaN	NaN
prev_state	NaN	NaN	NaN	NaN
prev_zipcode	99999	9.401325e+04	1220	13041.388096

Figure 4. Basic statistics - households dataset

Survey Vehicles:

Dimension: 79,011 rows x 38 columns

Description:

This dataset contains detailed vehicle information such as power, fuel, body type, model, and year of manufacture. It also contains our target variable, `veh_type`, which we are interested in predicting in the classification.

Basic statistics:

	max	mean	min	std
body_type	99	3.548094	1	7.495593
body_type_other	NaN	NaN	NaN	NaN
cylinders	99	17.248173	1	33.648750
cylinders_other	NaN	NaN	NaN	NaN
epa_id	37370	12828.417964	8	7202.126501
fuel_type1	9	1.053498	1	0.371608
fuel_type2	7	3.045369	1	0.817553
insurance	9	2.726694	1	2.264517
model_year	9999	2244.641081	1931	1370.393925
ownership	9	1.061782	1	0.448937
ownership_other	NaN	NaN	NaN	NaN
power_train	9	2.719549	1	2.443190
power_train_other	NaN	NaN	NaN	NaN
purchase_type	9	1.468137	1	0.705022
transmission	9	1.268464	1	0.913655
veh_make	99	33.449334	11	15.915359
veh_make_other	NaN	NaN	NaN	NaN
veh_model	REDACTED	NaN	REDACTED	NaN
veh_type	9	1.999657	1	0.520389
vehicle_used	2	1.339258	1	0.473471
why_no_travel	99	11.171496	1	28.100918
why_no_travel_other	NaN	NaN	NaN	NaN

Figure 5. Basic statistics - vehicles dataset

Feature Engineering

The next step is the actual manipulation of the appropriate features. As seen from the list above, there are over a hundred features within these three datasets from which the model will train. However, prior to the creation of the model, feature selection must be done in order to eliminate some features. In the current analysis, it has been discovered that such features as `'race', 'race1', 'race2', and 'race3'`, among others, are collinear, and thus a few can be removed at random prior to future analysis.

This project will explore specific relationships across features that relate to the final response variable: vehicle type. More specifically, in the preliminary exploratory data analysis, specific features that have been identified to be relevant include `'empl_occupation', 'citizen', 'race1', 'age', 'school_home', 'gender', 'person_trips'`, among others. However, each feature's true relationship to the response variable will be determined by doing ex-post inference analysis/iterating

using several machine learning models. Feature scaling, using general standardization and normalization techniques, will also be conducted. By disciplining these features, the accuracy and generalizability of the model will be improved.

For such a large dataset, feature engineering will be both comprehensive and revealing. Feature engineering, feature selection, model creation and the derived inference will proceed as an iterative process.

Timeline of Models:

Step 1:

Linear Models: Linear Regression, Logistic Regression

Nonlinear Discriminative Models: SVM with a nonlinear kernel, Decision Trees, Random Forests

Ensemble Methods: Boosting, Bagging

Step 2:

Nonlinear Predictive Models: Neural Networks, and (if time permits) Deep Learning models using TensorFlow, Keras, and/or an exploratory visualization using Word2Vec

Performance Analysis

The final step will be a performance analysis of the models and eventual selection of the optimal model. On the prediction side, this will entail splitting into training, development, and test data, cross-validation, and running the optimal model on the test data.

On the inference side, the goal is to find the features and/or combinations of features that ultimately have the most impact on the response variable.

Inference and Ex-Post Analysis:

We will analyze the relationship between dependent variables in our linear models and the target variable of vehicle type. In our discriminative models, we will visualize the decision boundaries and make inferences based on, for example, decision tree paths.

If time permits, we will present some methods from Susan Athey's research of ML for Inference and Causal Effects. As with other problems in the social sciences, our project has a need for both ML prediction and after-the-fact interpretation.

Pitfalls

Imbalanced Data

In the ongoing data analysis, it has already been found that much of the response variable's data are skewed toward gas-powered automobiles, significantly more than any other type of vehicle, including hybrids, EVs, bicycles, and public transit, as shown in Figure 6.

This imbalance will be addressed by upsampling the data. The effect on error rate has yet to be determined, but the expectation of the upsampling is that other performance metrics, such as the Kappa statistic, will be improved.

Other possible performance metrics to use for imbalanced data would be the F-score. Since F1 score is a measure of both precision and recall, we will drive the feature set development based on how well our classifier is performing on the positively classified data.

Since our data are unbalanced, we used a baseline of predicting all instances as gasoline cars:

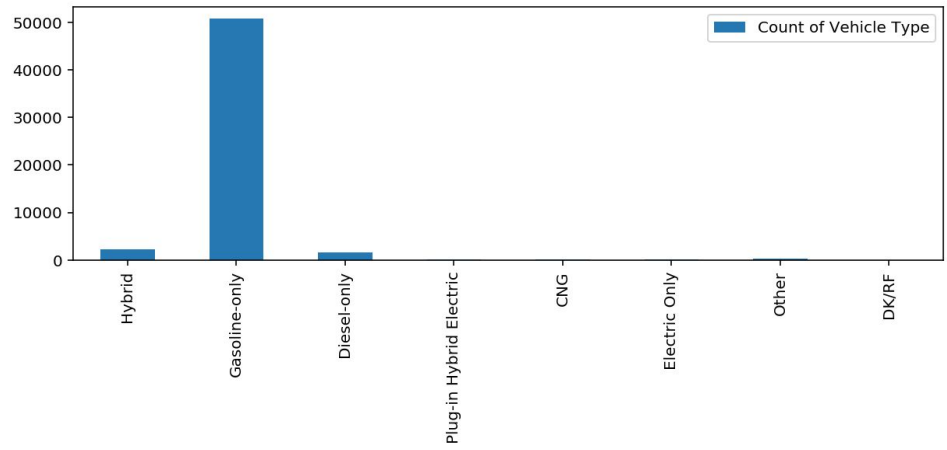


Figure 6. Imbalanced data in vehicle type.

Predict all gasoline cars as the baseline.

```
baseline_predictions = [2 for x in range(len(test_labels))]
print('Baseline Accuracy: {}'.format(accuracy_score(baseline_predictions, test_labels)))
print('Baseline F1-Score: {}'.format(f1_score(baseline_predictions, test_labels, average='weighte
d')))
```

```
Baseline Accuracy: 0.9197181910226122
Baseline F1-Score: 0.9581804197340952
```

Next, we ran SVM, KNN, Logistic Regression, and Decision Trees, and compared their accuracies to the baseline model. We used features such as education, gender, income, number of students per household, number of workers per household, residence type, ownership type, trip count, bike count, home zip code, and many more! We used feature sets ranging from 51 features to hundreds of features.

Prediction using Support Vector Machine (SVM):

In [26]:
```
# Adding the necessary features to the model
X=new_joined_df[['education','gender']]
# Y is the target variable
Y=new_joined_df['fuel_type1']
# Dividing the data into train and test ,to check the accuracy of the model
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, random_state=42)
# we create an instance of SVM to fit our data
clf = svm.SVC()
# Fitting the train data to predict the target for the test data
print("Training the model..........")
clf.fit(X_train, y_train)
# Predicting the target based on the training done
print("Pricting based on the training on the model..........")
pred=clf.predict(X_test)
print(clf.predict(X_test))
# Measuring the Accuracy Score of the Data
print("Accuracy Score")
print(accuracy_score(y_test, pred))
```

It's outputted accuracy is 95.74%, which is 4% better than the baseline.

Prediction using kNN, using k=5:

The feature set for the model below consists of all the features joined across households, persons, places, and vehicles in the CHTS survey tables. This has 324 features in total after performing some preprocessing.

```
#initialize knn classifier model (k = 5)
knn = KNeighborsClassifier(n_neighbors=5)

# fit the model
knn.fit(X_train, y_train)

# predict the target values
pred = knn.predict(X_test)

# evaluate its accuracy
print(accuracy_score(y_test, pred))
```

```
0.938864628821
```

It's accuracy is 93.9%, which also greater than the baseline.

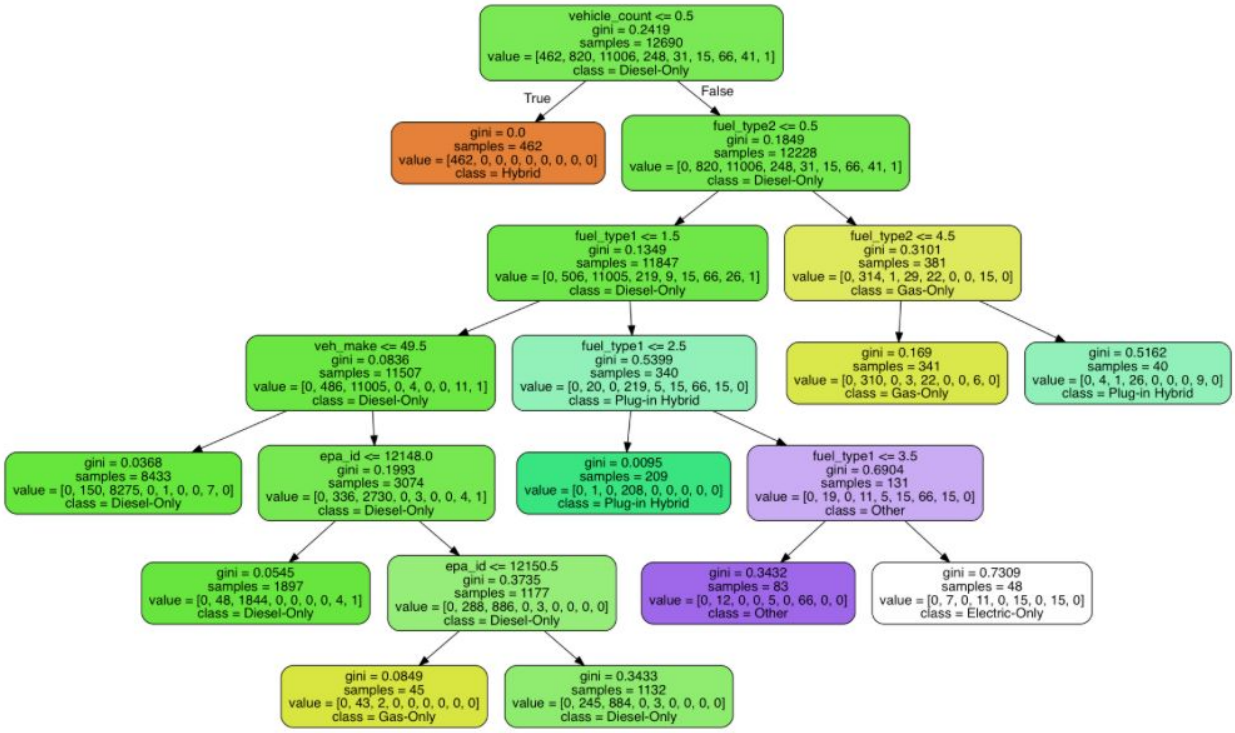
Prediction using Logistic Regression:

Using the same featureset as the decision tree analysis, we ran a logistic regression and got an accuracy of 92.5% and an F1 score of 95.2%. So this only slightly better accuracy than a baseline reading.

Prediction using Decision Trees:

Next we ran a Decision Tree on a smaller Feature set of 51 quantitative variables. Setting the maximum leaf node hyperparameter to 10, we got an accuracy of 96.2% and an F1 score of 97.0%. So the accuracy improved by 4.3% and the F1 score increased by 1%.

The visualization of the resulting decision tree is as follows:



The first split leads to an interesting insight. It splits on vehicle count where the split point is 0.5. So if a household has 3 people but only one car, then it is 100% likely according to this model, that the one car will be a hybrid. This is because the average vehicle_count value for the house would be 0.33, which is less than 0.5.

More feature engineering and preprocessing is needed to get more interpretable splits further down the tree.

Project Timeline

Round	Phase	Description	Duration
1	Data Cleaning	Combine data from different sources and perform fundamental cleaning (Owners: everybody) Descriptive Statistics (Owners: everybody)	Oct. 30 - Nov 2nd, 2017 **Completed**

		1: box + whisker plot (Vik) >1: correlation matrix (Pavan) - For each dataset: - Household, places, persons, vehicles - Use these insights as we move forward with feature engineering >1: stats (means, std, etc) (Pavan) 1: data types (Pavan) 1: description of categorical vars (Vik)	
2	First Round Algorithm Prediction	Predict with several ML algs: Try Linear Models (KNN and Logistic Regression), Categorical Predictions (SVM, Decision Trees). - KNN (Pavan) - Decision Trees (Vik) - Logistic Regression (Shrestha) - SVM (Udit) Include a visualization of results. We provide snippets of the 1st iteration of Rounds 1 and 2 as part of this proposal.	Nov 2nd - Nov 5, 2017 **Completed 1st iteration**
2.5	Feature Engineering	Coming together and creating a giant feature set	Nov. 7
3	Model Tuning and Feature Selection	Tune the model based on some chosen performance metric (RMSE, Kappa Statistic, F-Score) Owner: (Feature selection: PCA (Principle Component Analysis): Shrestha Lowest Variance: Udit Chi-Square or other: Vik Model selection: Kappa Stat CV: Pavan	Nov. 6 - Nov. 7

		)	
4	Inference and Ex-post Analysis	Interpret your models using some form of impact analysis in an ML setting Step 1: use regression model, decision tree paths to come up with insights about the independent and target variables. Regression analysis: Shrestha Decision Tree: Vik SVM: Udit, research online how to interpret Step 2: Use a Confusion matrix to identify what you got wrong in the classification. Use this analysis to iterate on feature engineering. CF Matrix: Pavan Additional features: everybody Step 3: Possibly add research from Susan Athey's work in ML for Inference (if time permits) Owners: (everybody, per model assigned above)	Nov. 6 - Nov 15
5	Nonlinear Methods	Neural Networks, Deep Learning (if time permits) Owners - 3 people: (Vik, Pavan, Shrestha)	Nov 16 - Nov 20
6	Ensemble Methods	Ensemble methods to optimize ML algorithms: Owners - 2 people: (Udit, Shrestha)	Nov 16 - Nov 17
7	Iterative Development	Iterate on rounds 2 through 6. Owners: (everybody)	Nov 17 - Nov 24

8	New Ideas / Finishing Touches	Trying out new ideas based on any insights gained from the process. Finalize the pipeline. Optional idea - Unsupervised clustering	Nov 25 - Nov 28

Summary

The goal of the project is to determine the vehicle type based on personal attributes, creating a prediction model as well as a causal inference model to understand which attributes have the most impact. This information can be used in both policy and corporate settings, as understanding the types of people who drive certain vehicles can lead to specific tax incentives as well as more effective targeted advertising campaigns, respectively.

The methods described above will be comprehensively tested to determine not only the best prediction model but also a robust causal inference model. Our time will be iteratively spent in feature engineering, determining the relevance and impact of the appropriate features for future training into the optimal machine learning model. This model will be chosen based on an extensive cross-validation of the methods over several different metrics, including accuracy, F-score, and Kappa statistic.