Machine Learning Engineer Nanodegree
Capstone Proposal

Salvador Núñez
*December 18, 2017*

## Proposal

### Domain Background
Advancements in the electric automobile industry are contributing to the disruption of the energy industry, particularly for electric utilities. Electric vehicle (EV) sales are expected to increase by 117%[1] in 2017. Countries like Norway, Netherlands, and China[2] plan to put many millions of EVs on the road by 2020. Growing adoption of EV means that electricity consumption for EV owners could cause dramatic and unpredictable shifts in electricity demand. Electric utilities are gathering and analyzing data on how the growing popularity of EVs could potentially stress the electric grid's infrastructure and develop a better-informed investment strategy as a result. For example, Con Ed's SmartCharge New York program offered to *pay* customers 5 cents per kWh[3] to charge during off-peak hours. While energy curtailment and time-of-use rates (TOU) can partially support the economics of this program, the real value comes from the data the program will collect. The analysis performed on these data may involve nonintrusive appliance load monitoring (NIALM) to disaggregate EV loads from other appliances and therefore detect EV owners. For example, the electric load of a household can be separated into different appliances through harmonic feature analysis and multiple-class support vector machines.[4] On the other hand, usage disaggregation has also been achieved through unsupervised machine learning techniques using hidden Markov modeling (HMM) and residual analysis.[5]

### Motivation
I´m an EV owner who is passionate about data and sustainability[6] and has professional experience in the cleantech and smartgrid sectors. I'm also about to install a level 2 EV charging station at home and I'm keen to understand how it will affect my electricity bill.

### Problem Statement
This project will be approached as a supervised learning classification problem. The goal is to analyze Advanced Metering Infrastructure (AMI) data to classify which residences have electric vehicles. Furthermore, the goal is also to predict when, or classify time intervals, when EVs are being charged. The provided only inputs are a training dataset of house ids with 30-minute energy readings in kWh, together with a labeled training dataset which denotes which houses are charging EVs during which time intervals with a 1 or a 0.

---

[1] "EV's Charging Up: Sales Beating Hybrid Into 2017 Expected to Be a Record Year". *Consumer Federation of America,* April 12, 2017. https://consumerfed.org/press_release/evs-charging-sales-beating-hybrid-intro-2017-expected-record-year/

[2] Hockenos, Paul. "Norway spearheads Europe's electric vehicle surge". *EURACTIV,* March 23, 2017. http://www.euractiv.com/section/electric-cars/news/norway-spearheads-europes-electric-vehicle-surge/

[3] Walton, Robert. "New ConEd EV program to reward customers for off-peak charging" *Utility Dive,* April 18, 2017. https://www.utilitydive.com/news/new-coned-ev-program-to-reward-customers-for-off-peak-charging/440639/

[4] Jiang et al. "An Approach of Household Power Appliance Monitoring Based on Machine Learning". Intelligent Computation Technology and Automation (ICICTA) – IEEE. Jan 14,2012.

[5] Pattem, S. "Unsupervised Disaggregation for Non-intrusive Load Monitoring ". Machine Learning and Applications (ICMLA) –IEEE. Dec 12,2012.

[6] Núñez, Salvador. "5 Lessons in Data and Sustainability". MEng Alumni Magazine – Fung Institute for Engineering Leadership UC Berkeley, May 2017.

## Datasets and Inputs

A labeled training dataset consisting of 1,590 rows and 2881 columns has been downloaded from GridCure's website[7]. The data contains one categorical variable for the 1590 "House IDs" and 2880 continuous variables consisting of two months of energy (kWh) data taken at 30 minute intervals (i.e. 2x24x60 + 1 =2,881). Similarly, there is a separate dataset also containing 1,590 rows and 2881, corresponding to the training data labels. Instead of containing energy values in kWh, each row contains binary labels where 0 corresponds to no EV charging and 1 corresponds to EV charging for that particular "House ID". This format is inconvenient for modeling since there are several labels per row.

Preliminary inspection of the data indicates that about 30% of the houses have EV´s charging for at least one 30-minute interval. Furthermore, on average these houses charge EVs about 8% of the time. This means that the classes are not balanced. A randomized train_test_split will be performed on the 1,590 houses and the datasets will have to be transformed into a dataframe with 1 target label per row.

## Solution Statement

The proposed solution is to apply my domain knowledge about the industry, unsupervised machine learning techniques, and supervised learning techniques, to classify which time intervals, and therefore which houses, are charging an electric vehicle. This will be achieved by defining custom data transformations to extract or engineer new features, and subsequently training a classifier that generalizes and can be applied to other similar datasets.

## Benchmark Model

In US Patent US 9,576,245 B2[8], Fischer et al. describes a method for identifying EV owners. The machine-learning model predicted EV ownership status after defining features in the AMI data such as increases/decreases in power load by certain pre-determined amounts (e.g. 1-2 kWh), a particular frequency of such increases/decreases, and temporal spacing between such events. Understanding the overall demand curve of residential energy was also an important consideration since large loads during off-peak hours could give a clearer signal of EV ownership. These signals get further strengthened by incorporating other datasets, such as additional monitoring services in the household, information on the appliances in the household, the size of the home, the geographical location of the home, weather data, etc. According to Utlity Dive[9], Opower, Inc, the company owning the patent, processed over 40% of all residential energy consumption data and nearly two-thirds of AMI data in the US. Opower had extracted other features from these AMI data that could be fed into their models, such as usage disaggregation[10] or load curve archetypes[11] derived from unsupervised learning techniques like k-means clustering. Finally, engaging with these residential energy customers with personalized communications allowed Opower to obtain responses to EV rebate programs that could be used to label new data and improve the model. The scope of this project is limited due the lack of access to many of these features. Furthermore, the bias-variance tradeoffs in training the model will be quite different from the one in the patent since the dataset under consideration is many orders of magnitude smaller. In any case, the patent does not include information on obtained accuracy of the described model. Therefore, the simplest benchmark is used on this project: a naïve predictor assumes that there are no EV owners in the dataset.

---

[7] Optional Predictive Modeling Challenge. GridCure, https://www.gridcure.com/contact/

[8] Fischer, et al. Identifying Electric Vehicle Owners. United States Patent US 9.576,245 B2. United States Patent and Trademark Office. Feb. 21, 2017.

[9] Walton, Robert. "Ontario power providers tap Opower for efficiency, DSM offerings" *Utility Dive,* Jan 7, 2017. https://www.utilitydive.com/news/ontario-power-providers-tap-opower-for-efficiency-dsm-offerings/411688/

[10] Fischer, Barry. "This neat data algorithm unlocks the power of smart grid technology—without using smart meters" *Opower, Inc.,* July 29, 2014. https://blogs.oracle.com/utilities/data-algorithm-smart-grid-without-smart-meters

[11] Fischer, Barry. "We plotted 812,000 energy usage curves on top of each other. This is the powerful insight we discovered." *Opower, Inc.* October 13, 2014. https://blogs.oracle.com/utilities/load-curve-archetypes

## Evaluation Metrics

An unlabeled test dataset is also available for download. Only GridCure has the ability to score the model on the test data set. Unlike Kaggle competitions, in this case there is no automated way of submitting predictions and obtaining a test score. Therefore, a validation dataset will be separated from the modeling exercise, and this project will be evaluated solely on the F1 score of the trained model on the validation dataset. Additionally, a confusion matrix to summarize precision and recall will be calculated for both questions under consideration: a) whether a "House ID" owns an EV and b) whether there is EV charging during a 30-min time interval.

## Project Design

First, 25% of the data will be removed from data exploration and model training and only used as the validation set. This process will be done through random assignment, while ensuring that there is the same proportion of EV owners in both training and validation sets. Next a series of transformations will be performed to extract or engineer features from the hidden time dimension. Practically speaking, the raw wide dataset (1590x2881) will pivot into a long dataset that contains one row per house-interval combination (4,579,200x#columns). This will yield the traditional format for supervised learning having a matrix **X** and a target vector **y**.

Raw Data

| HouseID | Interval_1 | Interval_2 | Interval_3 | Interval_4 | ... | Interval_2880 |
|---|---|---|---|---|---|---|
| 11655099 | 0.95 | 0.826 | 0.361 | 0.238 | ... | 0.728 |
| 11633257 | 0.353 | 0.327 | 0.358 | 0.292 | ... | 0.289 |
| 11651552 | 0.15 | 0.181 | 0.15 | 0.15 | ... | 0.113 |
| 11636092 | 2.088 | 2.075 | 2.121 | 2.098 | ... | 0.34 |
| 11647239 | 1.416 | 1.25 | 1.27 | 1.258 | ... | 1.571 |

Raw Label

| HouseID | Interval_1 | Interval_2 | Interval_3 | Interval_4 | ... | Interval_2880 |
|---|---|---|---|---|---|---|
| 11655099 | 0 | 0 | 0 | 0 | ... | 0 |
| 11633257 | 0 | 0 | 1 | 0 | ... | 0 |
| 11651552 | 0 | 0 | 0 | 0 | ... | 0 |
| 11636092 | 1 | 0 | 0 | 0 | ... | 0 |
| 11647239 | 0 | 0 | 0 | 0 | ... | 1 |

Transformed Data with Label

| HouseID | Interval | Day | ClusterShape | OffPeakHours | Energy(kWh) | Label |
|---|---|---|---|---|---|---|
| 11655099 | 1 | 1 | 3 | 0 | 0.95 | 0 |
| 11655099 | 2 | 1 | 3 | 0 | 0.826 | 0 |
| 11655099 | 3 | 1 | 3 | 0 | 0.361 | 0 |
| ... | ... | ... | ... | ... | ... | ... |
| 11647239 | 2880 | 60 | 4 | 1 | 1.571 | 1 |

These new features may include: the day of the interval (continuous), the associated load-curve by applying k-means on the daily load curves (categorical), a binary estimate of whether the interval is during peak hours (categorical), an estimate of the incremental load for the house during that interval (continuous), etc.

Once the new features are engineering and the data is transformed, I will train a classifier on the dataset and optimize it to obtain the greatest cross-validation score. I am unsure what kind of classifier might perform best, but plan on trying neural networks, support vector machines, boosted decision trees, and logistic regression. It is likely that some of the classifiers, like neural networks, may overfit the data, so I also anticipate needing to apply some level of regularization.