Machine Learning Engineer Nanodegree

Capstone Proposal

Salvador Núñez

December 17, 2017

Proposal

Domain Background

Advancements in the electric automobile industry are contributing to the disruption of the energy industry, particularly for electric utilities. Electric vehicles (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 demand.

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. Analysis on these data can involve disaggregation of EV consumption from other types of energy consumption and discriminating between EV owners and non-EV owners using Advanced Metering Infrastructure (AMI) data.

Motivation

I´m an EV owner who is passionate about data and sustainability[4] 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

The goal of this project is to analyze AMI data to predict which residences have electric vehicles. Furthermore, the goal is also to predict when, or during which intervals, the electric vehicles are being charged.

---

[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] 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[5]. The data contains a "House ID" and two months of energy (kWh) data taken at 30 minute intervals for 1590 houses.

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, it contains binary labels where 0 corresponds to no EV charging and 1 corresponds to EV charging.

Solution Statement

The proposed solution is to transform the provided dataset to reveal or engineer additional features, train a neural network classifier, and test the accuracy on a validation dataset which will be excluded from the training.

Benchmark Model

In US Patent US 9,576,245 B2[6], 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. Heating, ventilation, and air conditioning (HVAC) energy consumption is typically the largest load in the household and is typically seen during peak hours. 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[7], 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[8] or load curve archetypes[9] 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 training data and improve the model.

---

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

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

[7] 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/

[8] 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

[9] 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

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.

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 prediction accuracy 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 the customer is an EV owner and b) whether the 30-min time interval is associated with EV charging.

Project Design

First, 25% of the data will be removed from data exploration and model training and only used as the validation set.

Then, before performing supervised learning, I will use unsupervised learning and my domain knowledge to explicitly add features hidden in the data. For example, I will reshape the data to have 1 record per day and apply k-means to understand if there are some common daily load curve archetypes in the data. I may also summarize some statistics for each household, such their average daily energy use. This will allow me to compare each households daily energy use with the mean daily energy use of the sample. Ultimately, instead of having a very wide dataset, I will transform this dataset into a very long one with multiple engineered features from the dataset, including the kWh reading, and a single label for each row. This differs drastically from the current dataset where the training dataset is very wide and has multiple labels per row. These transformations will need to be applied to the validation dataset as well prior to feeding it to the neural network, but they should not inform the design of these features.

Once the new features are engineering and the data is transformed, I will run a neural network on the dataset and optimize it to obtain the greatest cross-validation score. If the neural network seems to overfit the data. I may try to apply some regularization or compare the performance with other models such as support vector machines (SVM), boosted decision trees, and logistic regression.