## Machine Learning Engineer Nanodegree

## Capstone Project Report

## Salvador Núñez

## February 24, 2018

# Definition

## Project Overview

Advancements in the electric automobile industry are contributing to the disruption of the energy industry. Electric utilities are increasingly interested in identifying the load from electric vehicles (EVs) in order to develop a better informed investment strategy and create incentives for residential customers to charge their EVs during certain times in the day. This project analyzes the electricity consumption of houses to identify a) if there is an electric vehicle (EV) charging at each house and b) during which time intervals the EV is charging at the house. A [dataset](https://www.gridcure.com/wp2/wp-content/uploads/2016/05/EV_files-2.zip) consisting of 1,590 houses has been downloaded from GridCure’s [website](https://www.gridcure.com/)[[1]](#footnote-1), containing labels to denote which houses are charging their EVs and when. The website also contains a separate unlabeled test dataset with an additional 699 houses. The goal of this project is to create a series of data transformations and machine learning models that correctly predict which houses have EVs and when the EVs are charging. The performance of the models are benchmarked against a validation data set separated from the original labeled data set. Finally, these transformations and model are used to submit predictions for the 699 test houses.

## Problem Statement

This project will be approached as a supervised learning classification problem. The first strategy used to solve this project is to find ways to decrease the imbalance of the classes. While 31% of the houses in the training dataset are labeled at charging during at least 1 time interval, that corresponds to only 2% of the intervals or data points yielding a positive label. Training a model to accurately predict a positive label when it only occurs 2% of the time in the training data set will be very difficult. To help with this, first houses that have EVs are predicted. Then, using only that subset, time intervals when EVs are charging are predicted. By doing so, the prevalence of a positive label is increased from 2% to 8%. This four-fold increase in a positive label will make the training much easier.

The second strategy used is to augment the provided training data with implicit temporal information. 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). However, we know that means there are 60 24 hour cycles in the data, but those cycles are not represented in the dataset. Therefore, the day and the time of day corresponding to each interval are explicitly added to the data.

Using the daily cycles in energy consumption, consumption archetypes can be defined. Companies like Opower Inc, have developed [load curve archetypes](https://blogs.oracle.com/utilities/load-curve-archetypes)[[2]](#footnote-2) derived from unsupervised learning techniques like k-means clustering. Load curve archetypes are also applied here with the objective of creating clusters that effectively discriminate between EV houses and non-EV houses.

Finally, Gradient Boosting Decision Trees (GBDT) to classify the houses and the time intervals. A GBDT model was applied since they [often perform well](https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/)[[3]](#footnote-3) on imbalanced datasets.

## Metrics

The performance of the model(s) developed for this project will be evaluated based on the F1 score of the trained model on the validation dataset. The F1 score is a better evaluation metric than accuracy when considering a model trained on uneven class distribution. The F1 score is calculated as the harmonic mean of precision and recall.



{\displaystyle F\_{1}=2\cdot {\frac {1}{{\tfrac {1}{\mathrm {recall} }}+{\tfrac {1}{\mathrm {precision} }}}}=2\cdot {\frac {\mathrm {precision} \cdot \mathrm {recall} }{\mathrm {precision} +\mathrm {recall} }}}



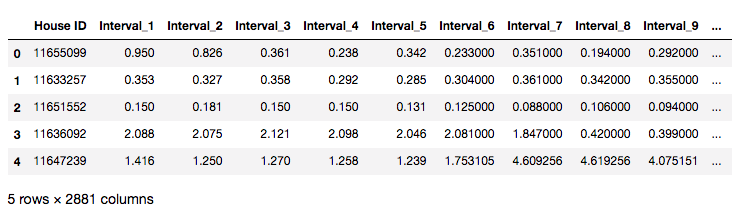


High recall attempts to minimize false negatives, whereas precision tries to minimize false positives. Accuracy works best if false positives and false negatives have similar cost. In this case, a false negative (incorrectly predicting a house or time interval is not charging an EV) is much more expensive than a false positive (incorrectly predicting that a house or time interval is charging an EV, when it isn’t).

# Analysis and Methodology

## Data Exploration

The “wide” data format for the data set does not conform to the traditional format for scikit learn with a matrix **X** and a target vector **y** corresponding to exactly 1 prediction per row. It also contains 4 houses missing between 48 and 144 data points.



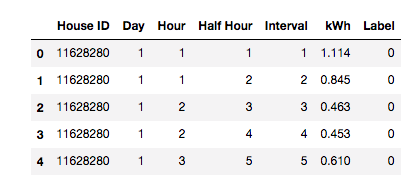
## Exploratory Visualizations

Preliminary exploration of the raw dataset shows a long tail distribution of the energy consumption of the 30-minute interval, as well as the distribution of energy consumption for houses over the entire 2 months. This indicates that there may be some outliers in the data, but the data doesn’t seem to be multimodal.

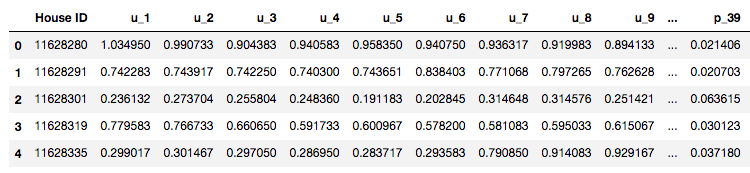
## Data Preprocessing

All training data set and the training labels are combined and pivoted to “long” format. Then, the implicit temporal dimensions are explicitly added with the assumption that sequence in the interval column names can be translated to 60 cycles of 48 30-minute intervals. With this temporal information, the missing values are filled for the 4 houses missing data points. These are filled with the average energy value for that house during that time of day. This results in the following data set without any null values:



The data is later aggregated at the house level and new labels are created at the house level. If a house has **any** time interval positively labeled for charging EVs, it is a positively labeled “EV house”. If not, it is a “non-EV house”. These house aggregations are further described with the mean and standard deviation of the kWh the house consumes every half hour. The means are also divided by the average total daily consumption to produce a metric for the average percent of daily energy spent in that half hour. This last summary statistic, the percent of daily energy spent, is particularly helpful in isolating the load shape of the house, irrespective of the total amount of energy the house may consume. These columns are named with the following convention, where “i” is an interval between 1 and 48, corresponding to all the half hour intervals in a day:

* u\_i – mean during the ith interval for that house
* s\_i – standard deviation value during the ith interval for that house
* p\_i – average percent of the daily consumption during the ith interval for that house

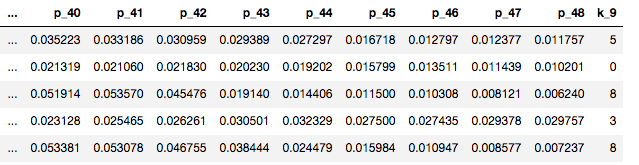


Before any additional processing, this data set aggregated at the house level is split between training and validation groups, using a test size of 0.25.

Then, an additional column is added to the separated training data set: a cluster assignment. The data set is clustered through k-means using only the columns describing the percent of daily consumption during the ith interval. The “optimal” k was selected based on the k that would maximize the ratio comparing the cluster with the highest percent of positive labels with the cluster with the lowest percent of negative labels, for each set of clusters produced by k ranging from 2 to 10.



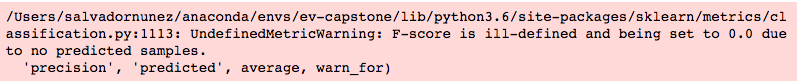
In this case, the optimal “k” was 9 and the column “k\_9” was added to the aforementioned dataset, assigning a cluster from 0 to 8 to each house.



Next, a Gradient Boosted Decistion Tree model (GBDT) is trained on this data set, resulting in a training F1 score of 0.98. However, this F1 score is on the training data set and has a fair degree of overfitting. Instead, an F1 score must be obtained from the validation data set and compared against the benchmark.

To do so, a k-nearest neighbors (KNN) algorithm is trained on the training data set and used to predict a cluster for the validation data set. This is because the validation data set was excluded from the k-means exercise in order keep the validation data set unexposed to any supervised or unsupervised training technique. Once column “k\_9” is appended to the validation dataset using KNN, the trained GBDT model is used to predict EV houses in the validation data set. This yields a validation F1 score of 0.75.

Benchmarking this score against a naïve predictor that assumes that there are no EV owners in the dataset did not make too much sense because this benchmark yields 0.00.



Therefore, two more naïve predictors were used: (a) a naïve predictor that sequentially alternates between predicting 0 and 1, and; (b) naïve predictor that predicts that assumes that all houses are EV houses. These scores were 0.367 and 0.487, respectively. The model does a much better job at predicting EV houses than these benchmarks.

## Implementation

## Algorithms and Techniques

## Benchmark

## Refinement

# Results

## Model Evaluation and Validation

## Justification

# Conclusion

## Free-form Visualization

## Reflection

## Improvement

1. Optional Predictive Modeling Challenge. GridCure, <https://www.gridcure.com/contact/> [↑](#footnote-ref-1)
2. 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> [↑](#footnote-ref-2)
3. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/> [↑](#footnote-ref-3)