**Electric Vehicle Detection with logistic regression and Support vector classification**

**Introduction**

This data set is provided by Gridcure as a practice problem. The data set contains training data, training data label and test data. The training data contains two months of smart meter power readings from 1590 houses taken at half-hour interval. Some of these houses have Electric vehicles charging and some do not. The training data label indicates the time intervals on which an electric vehicle is charging, with 1 denoting the presence of Electric vehicle and 0 denoting the absence of electric vehicle. The objective of this exercise is to determine:

1. Which house, given it’s smart meter power reading, has an electric vehicle?
2. When are the electric vehicle charging ?

**Methodology**

Major portion of this exercise is concentrated on identifying the household with EV given their smart meter reading. Achieving this objective is essentially a binomial classification problem. I employed following techniques to solve the problem.

1. Logistic regression (M1)
2. Support Vector Classification with rbf kernel (M2)
3. Support Vector Classification with polynomial kernel(M3)
4. Support Vector Classification with features of meter reading(M4)

Second objective was also approached as a binary classification problem. A logistic model was trained with time series feature and provided labels (M4). This model was then used to predict the intervals at which the EV charging was most likely to take place.

**Data Wrangling**

The data was provided in csv format. It was imported in jupyter notebook with pandas. For the first objective, Minimal data wrangling was done. The index of data was reset to House ID. The columns of data were interval number. There were total 1590 columns, one for each house and 2880 interval. Four households had null values for the data. They were simply removed and no imputation of any kind was done. For the second objective the data was transposed so that the rows were interval number and columns were House ID.

**Modeling**

Logistic modeling (M1)

*‘EV\_logistic.ipynb’ in repo*

As a most basic modeling approach, logistic regression was used. The time intervals were used as the features of the household. The house ID in training data which had EV were identified and labeled as 1 and the ones without EV were labeled as zero. The count of House ID with EV was less than half of that without EV which suggested that the training data was unbalanced.

The data was split into 3 sets for cross validation. The sequential nature of data was preserved by using TimeSeriesSplit() module of scikit learn library. A pipeline containing scaler and logistic regression model. The data was scaled to interval of (0,1) with Standard Scaler(). The logistic regression model was instantiated with balanced class weights due to unbalanced nature of data. The ‘C’ and ‘gamma’ hyperparameters were tuned using grid search method. The model was scored across validation sets using F1score. F1 score is the harmonic mean of recall and precision. This was done since data points labeled as 0 outnumber the data points labeled as 1and default scoring metric accuracy would not measure the usefulness of model in a meaningful way.

Support vector classification on raw data (M2 and M3)

*‘EV\_SVM\_rbf.ipynb’ and ‘EV\_SVM\_poly’ in repo*

The smart meter data of each household was used as features for training the support vector classification (SVC) model with ‘rbf’ kernel. The data preparation for this model was done following same steps as the data preparation for the logistic model. A pipeline with standard scaler and model instantiation was done. The SVC was instantiated with rbf kernel and class weight were set to ‘balanced’. The ‘C’ and ‘coef0’ hyperparameters were tuned using grid search. The model was scored using F1 score instead of accuracy. Another SVC model with a 3rd degree polynomial function was also created (M3). All other configuration was kept same as before for this model.

Support vector classification with features of the smart meter data(M4)

*‘EV\_features.ipynb’ in repo*

For this model, the mean, median, standard deviation, inter-quantile range, normalized range and normalized inter quantile range of the meter reading were calculated across the time intervals for each House ID. These features were then used to train the model and predict the label. The model was instantiated with 3rd degree polynomial kernel. The model was configured as the SVC models described above.

Logistic model for predicting interval (M5)

*‘EV\_interval.ipynb’ in repo*

In order to achieve the second objective, the training data was transposed so that the House ID were the columns of the data and intervals were the rows. From this data set one House with EV was chosen. The features of this data was created by computing 24 hr lags. Six such features were created and the model was trained on these features to predict the label for the corresponding House ID. The label count show that just 4% of total 2880 interval have EV present i.e. label is 1. This makes the data highly unbalanced.

The lag features reduced the number of non-null training data points to 2640. This data was split in to three fold with Time Series Split to preserve the sequential nature of data. A pipeline consisting of standard scaler and logistic regression model instance was created. The model instance was chosen to have balanced class weights. Hypermeters ‘C’ and regularization were tuned with grid search method. Scoring of the model was done with F1 score for reasons described above.

**Result and Analysis**

The best cross validation F1 score by model M1 was 0.67 with corresponding recall of 0.58 and precision of 0.80. A log loss value of 0.07 was calculated for training sample prediction. This shows that the model has a good accuracy but it predicts high number of false negatives. In other words, the model has a tendency to predict absence of EV when the EV are actually present in a household. This shows the importance of choosing F1 score as scoring metric. The M2 model performed even poorly with F1 score of 0.62, recall of 0.53 and precision of 0.75. The M3 model performed similarly with F1 score of 0.63, recall of 0.51 and precision of 0.83. The M4 models outperformed these models with F1 score of 0.72, recall of recall of 0.80 and precision of 0.64. The M4 model has highest recall of these models and F1 score shows a good balance between recall and precision. This model predicted that 269 houses out of 697 in the test data had EV.

To predict the EV charging interval, the logistic model was used (M5). This model performed well with F1 score of 0.89, recall of 0.91 and precision of 0.88. The model also had good accuracy with log loss of 0.09 for insample prediction. The model predicted EV charging on 370 out of 2640 intervals.

It also must be noted that the model was trained on features of data from just one EV charging house. The expansion of training domain of model will probably result in better performing models. The plot of the meter reading and predicted labels show that the instances of charging correspond to spikes in meter reading.

**Conclusion**

The features of smart meter power reading were better predictors of EV than the raw data. This shows the strength of feature engineering and predictive modeling based on the features. With raw training data, the support vector classification performed as well as simpler logistic model.