ACOREANA | | Volume 17, Issue 1, 2016 | | DOI : 16.10089/ACO171-01 | | ISSN : 0874 - 0380

Technologies for Detecting Electric Vehicle Presence in Modern Transport Networks

Michal Dvořáka

Adam Havlíčeka

The George Washington University

ABSTRACT

The global popularization of electric vehicles (EVs) poses an opportunity for the construction of micro-grid and smart community within energy internet on competent the massive and concentrated energy switching and routing in the local environment. Smart EV charging is a promising solution to manage EV charging load that relies on an accurate prediction of EV charging demands. Evaluation of household EV charging demand is a primary factor for designing smart EV charging solutions on household and neighbourhood level. However, this subject has not been adequately discussed in the research community. In this paper, several widely used machine learning algorithms are adopted to forecast the household electric vehicle presence situation such as Logistic Regression, Naive Bayes, Random Forest and XGBoost Classifier. The performances of the algorithms are evaluated and compared. A time series model called ARIMA is also proposed, that performed successfully with forecasting the future electric usage of individual family and to judge whether the family needs smart charging solution in the future or not. By using the above machine learning model, we can have a general understanding of the advantage and disadvantage of each model and when should it be used.

Keywords: EV, Logistic Regression, Random Forest, XGBoost Classifier, ARIMA, smart charging

INTRODUCTION

Electric vehicles (EVs) have been around since the late 1800s. They were very popular and a number of EVs were sold until about 1918. For example, in the year 1900, about 4200 automobiles were on the road, out of which 38% were electric, 22% gasoline powered, and 40% steam. With the advancement of gasoline engines, low-cost gasoline, and the invention of electric starter for the internal combustion engines, the interest in EVs completely declined. In spite of it, some automotive companies continued to work on research and advancement of EV technologies by experimenting with different types of propulsion motors, energy storage systems, and also incorporating advanced power conversion technologies. Rajashekara, K. (2013) has pointed out that, in the last ten years, plug-in hybrid electric vehicles (PHEVs) are attracting increasing interest in North America and in other countries. A plug-in hybrid could be a series or a parallel hybrid with its battery restored to full charge by connecting it through a charger to an external electric power source as in an EV. This requires a relatively higher capacity battery compared with the typical state-of-the-art HEV batteries. Additionally, significant enhancements beyond typical full HEV powertrain configurations would be required to properly handle the increased thermal management system loading and other factors associated with plug-in HEV usage. With their larger battery capacity of ~5−15kWh, PHEV can be driven in pure EV mode for short distances. In hybrid mode, the battery works as an HEV battery for power assist. Thus, a PHEV battery needs energy and power performance, requiring shallow cycle durability similar to that in HEVs and deep cycle durability such as EV batteries. GM's Chevrolet Volt, Toyota Prius, and Ford C-Max Energi are some of the PHEVs that are readily available in the market. Other auto makers are also planning to commercialize PHEVs in the very near future.

Pecan Street has hit an important energy recently with the installation of an electric vehicle (EV) charger that can charge from and discharge to the grid-a bidirectional flow of energy known as Vehicle-to-Grid, or V2G. For the first time in Austin, and possible in Taxes, energy was transferred from electric an electric vehicle’s battery pack to the electric grid’s distribution feeder. With this great convenience, more and more families will prefer to buy an electric car in the near future, thus, deploying a machine learning product to gather participants survey information and then forecasting whether selected family has EV or not offers great benefit to related industries for EV sales or later deploying smart charge system. However, there are two possible results, one is that people with or without electric vehicles feel happy to tell the his/her EV situation. For this groups, we can easily take action. Another result is that people take a

conservative attitude and are reluctant to reveal too much privacy, but there is no denying that they are likely to join the purchase of electric vehicle cars in the future. And this the part of the population that we're targeting. Instead, answering the question of whether you have an electric car, they may answer only the most basic questions such as 'Do you spend time at home on Monday', 'what is your ethnicity', 'what is your education level', 'what is your totally annual income' etc. We can actually play the answers from these questions. But how to select the most effective features for modeling since there are bunch of information from the Pecan Street Dataport as well as in the 2017 survey data. Therefore, a long period of feature engineering work has been done to control data normally distribution and to some extend reduce the possibility of data imbalance.

In this paper, machine learning methods taking their advantages of handling multi-dimensional and multi-variety data in dynamic environment, are attempted to study various households background. Several widely used machine learning algorithms are utilized to predict the electric vehicle household presence situation, as well as whether they do need smart charging system to reduce their electrical usage in the future. Forecast results are evaluated and compared. Through analyzing the prediction result such as metrics table, feature importance, confusion matrix and Receiver operating characteristic (ROC), we could clearly understand the advantage and disadvantage of each model based on different situations.

LITERATURE REVIEW

In recent years, the market share of EVs becomes more pronounced along with the maturity of related techniques and with the active promotions from governments and companies. T. Trigg et al has described that this ongoing vehicle energy type transition is leading novel challenges for the traditional power industry and at the same time providing an opportunity for the construction of energy internet at micro-grid and smart community levels. One of the obvious challenges for traditional power industry is that, the massive concentrated EV charging demands are likely to disrupt or damage current grid operations through the amplified daily peaks in a short future (C. Bikcora et al). The upcoming hardware and software infrastructure solutions based on energy internet should be competent to handle the high-speed growing tendency of power demand related to EV.

For an ideal situation, most of households could somehow live close to the electric charging station. But in reality, because a lot of the families live in other states or village, it is less likely for them to purchase an electric vehicle. However, this group is currently not in my research area. However, the prediction of EV presence in a household is also required by other smart household solutions. Home energy

management system (HEMS) is an important component of smart home solutions to realize energy planning and scheduling in a household [10]. EV charging has become a critical component of the entire power consumption. It is required to incorporate EV charging prediction into HEMS [11], [12], in order to schedule and adjust a household's entire power micro-generation and consumption profile efficiently to realize household energy bi-directional transmission on demand as well as household energy dynamic balancing. For considering the data balance and in order to create an equal data which contains households without electric vehicles, I decided to use much normal features for my model which is also kind of avoiding my data overfitting problem.

In 2012, Pecan Street's research sparked the purchase of 80 electric vehicles within a half-square-mile area – that's the highest concentration of consumer electric vehicles in the world. Since then, they have collected one-minute interval data on home EV charging and publish it through their Dataport website, where it is used by leading researchers from around the world. The data has led to insights on behavioral economics, the elasticity of EV charging profiles, grid system impacts from dense concentrations of private EV adoption, and aggregated charge management. More than thousands of electric vehicle owner and driver took the time to participating into completing the 2017 survey from pecan street. Therefore, we have firsthand information. Because the database data is updated in real time, I have used the most precise survey result.

RESEARCH METHODOLOGY

Firstly, SQL has been used to do the first step table merging in the Pecan Street Dataport web server, then Python 3.6 and common packages (pandas, numpy, matplotlib) were used to process the second step merging task on the google drive Colaboratory notebook implemented under google driver authentication package. The reason why I prefer this notebook but not anaconda is that Colab could switch running mode from CPU to GPU which could largely improve my computational ability to my huge dataset. Also in this process, (geopands, pygeocoder) functions were used to deal with geographic feature such as 'City' and 'Sate' then I can visually look at the location but not distinguishing the latitude and longitude. After Data Preprocessing was finished which contains data cleaning and table merging, I moved to phase two. In the phase one, I manually create a traditional dataframe for baseline model, and another time series dataframe for ARIMA model. Respectively in the early stage of the establishment of the two models, I did feature engineering job such as Exploratory Data Analysis (EDA) which provided the core information gained in underlying patterns of the dataset itself through visualization of distributions and

variable relationships as well as getting feature importance in order to deleting the strong effective feature in order to avoid data over fitting. Lastly, baseline modeling was done using Logistic Regression, Naïve Bayes, Random Forest, XGBoost Classifier, to determine whether specific households have electric vehicle or not. And then time series modeling was done using Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA), to use this family's historic data to predict electricity consumption for the next period of the time. Finally, analysis and conclusions were drawn based on the results of all processes in the project.

DATA

Metadata, electric_vehicles, electricity_ehauge_hour, survey_2017_all_partcipants and weather dataset was acquired from Pecan Street Dataport. The raw dataset was very huge and completed and the data format was many records traced by time interval, and my selection is using hourly dataset. On the average, each family has more than 8000 hourly records so I manually picked 37 dataids as electric vehicle owner families and another 37 dataids for non-electric vehicle families. About the weather dataset, I have combined the other three variables (temperature, windspeed, humidity) and created a new feature called ssd representing the human comfort. Through dataset refinement in the Data Preprocessing and EDA phases, the final dataset contained 33 features and 663,602 samples. The resulting dataset focused solely on year the Texas, Colorado two states including cities such as Humble, Round Rock, Plano etc in 2017. Because this is a time series dataset and baseline model will easily remember each family's records which directly reduced the model learning ability, I have to manually split 70% (43 families) as training dataset and 30% (23 families) as testing dataset. A list of factors utilized, associated Python data types, and a brief variable description can be found in Appendix A

DATA ANALYSIS

The first column, dataid, is a generated identifier to mask the individual family, and was immediately dropped from the dataset for baseline model but still kept in the time series dataset for deployment. Country and state values were updated and organized uniformly, to condense values such as "Colorado" and "Texas" into the same category. Any NaN values were replaced by 0 or unfilled for survey type feature. On the other hand, existed values are replaced by 1, filled or other meaningful values, and eventually all done by one hot encoding for the purpose of fit the baseline models later. Visualizations including python plotly, lineplots, areaplots, box and whisker plots, barcharts, geographic plots, correlation, missing map, and LMplots were generated with matplotlib and seaborn Python packages. The

sklearn package provided support for all modeling requirements, to include fitting the training data, and predicting the outcomes of test data. The Logistic model was the first classification model used to fit, predict, and score the data. Additionally, the feature importances function within the model's attributes was ran to score all the features according to effect each had on the model's outcome. Then followed by Naïve Bayes, Random Forest and XGBoost. Lastly, a time series ARIMA model was developed to fit, predict, and score the feature variables provided by the family historical data to forecast the next period of future time intervals. The code Is published on: https://github.com/Monarch2018/ev_prediction

KEY FINDINGS

The EDA process provided several visual displays to gather a baseline understanding of the nature of the data being utilized. Boxplots unanimously demonstrated a tendency towards left skewness with several outliers extending to the right. Each variable reviewed also had a relatively close grouping of the inner quartile range (IQR) in comparison with the full range including outliers. Insight of the modeling part, I could basically categories one as time series model and another as baseline model.

1. Time Series Model

For time series model, I have basically gone through 7 processes included importing libraries and dataset, Family choose, Test Harness, Persistence, Data Analysis, ARIMA Models, Model Validation. Skip the first phase as I have mentioned above already, we could choose a family as our prediction target, so let us pich first family whose data id is 410. We could use Test Harness to define a validation set and deploy a method for model validation of the last step. Then we could get dataset 297 samples, validation 68 samples. Running the persistence prints the prediction and observation for each iteration of the test dataset. The code ends by printing the RMSE for the model. In this case, the persistence model achieved an RMSE of 12.701. This means that on average, the model was wrong by about 12 electrical usage point for each prediction made. In data analysis phase, I did summary analysis and found that the large spread in series will likely make highly accurate predictions difficult if it is caused by random fluctuation (e.g. not systematic). In order to see the data distribution, I plot the Area plot, Histgram and Box & Whisker plot.

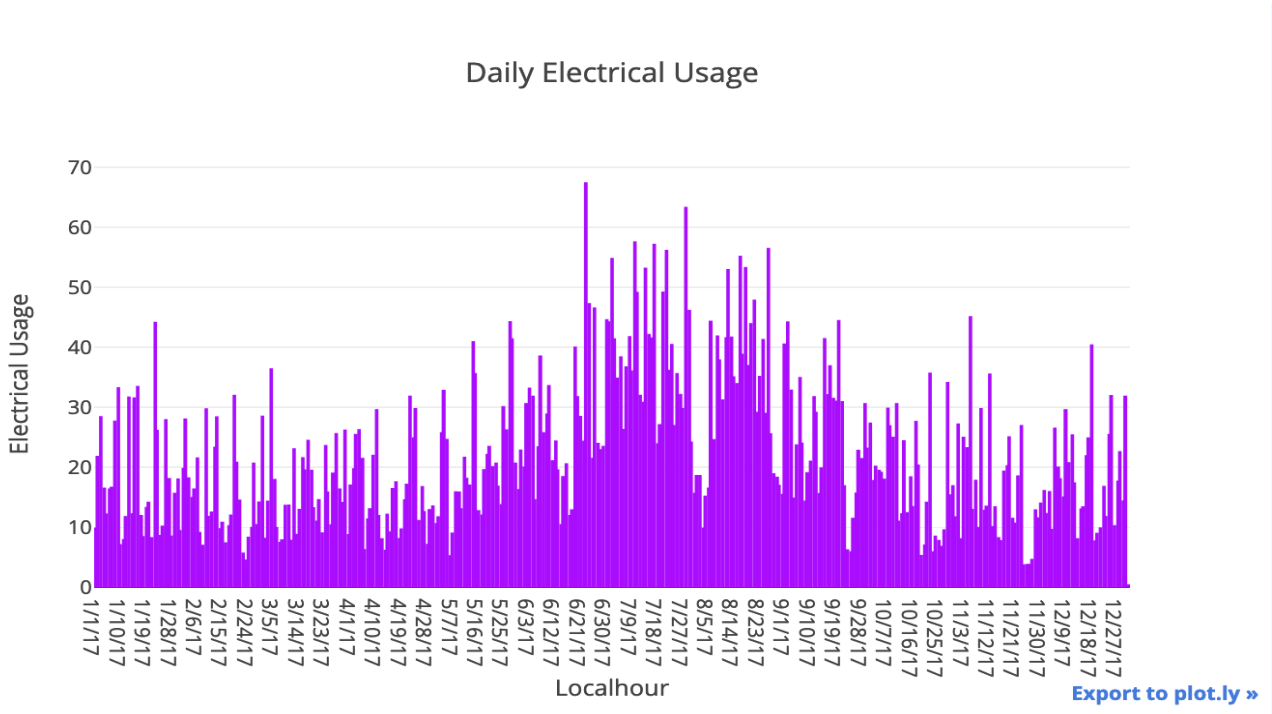

Figure 1. Daily electrical usage.

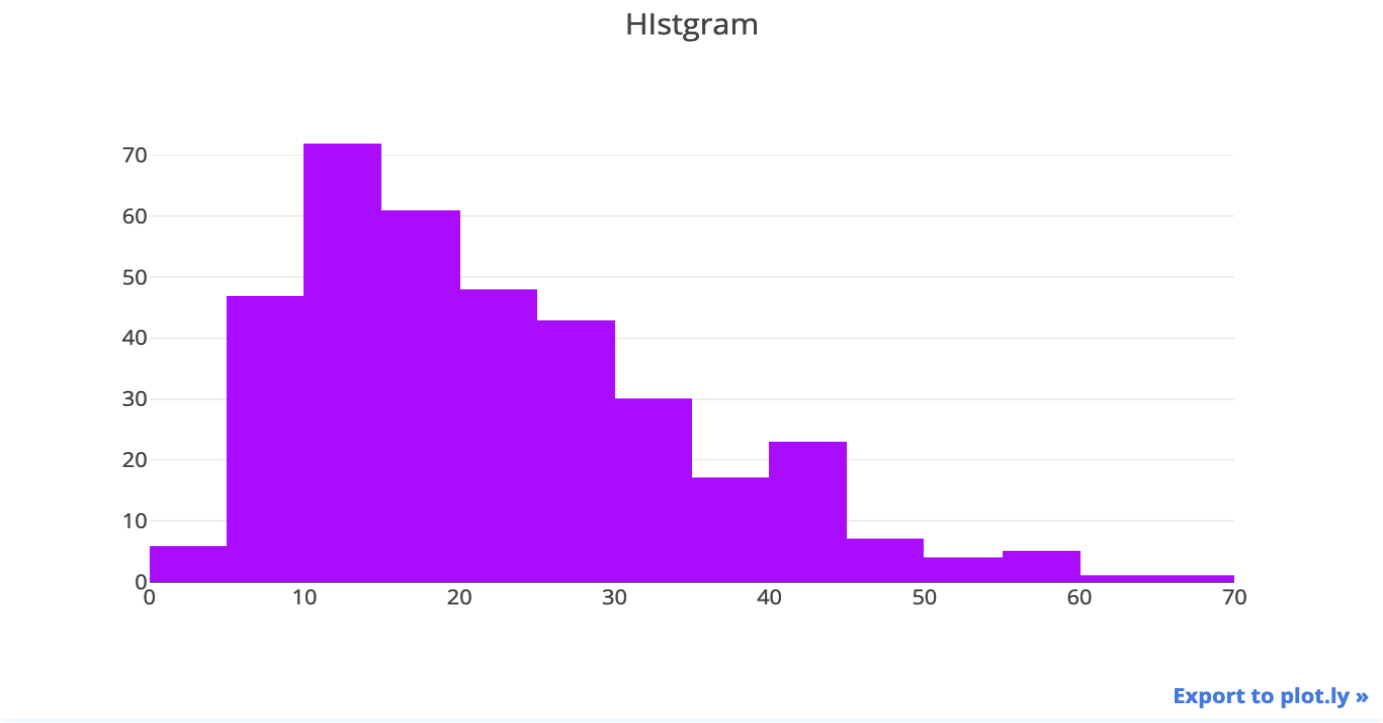

Figure 2. Series data distribution.

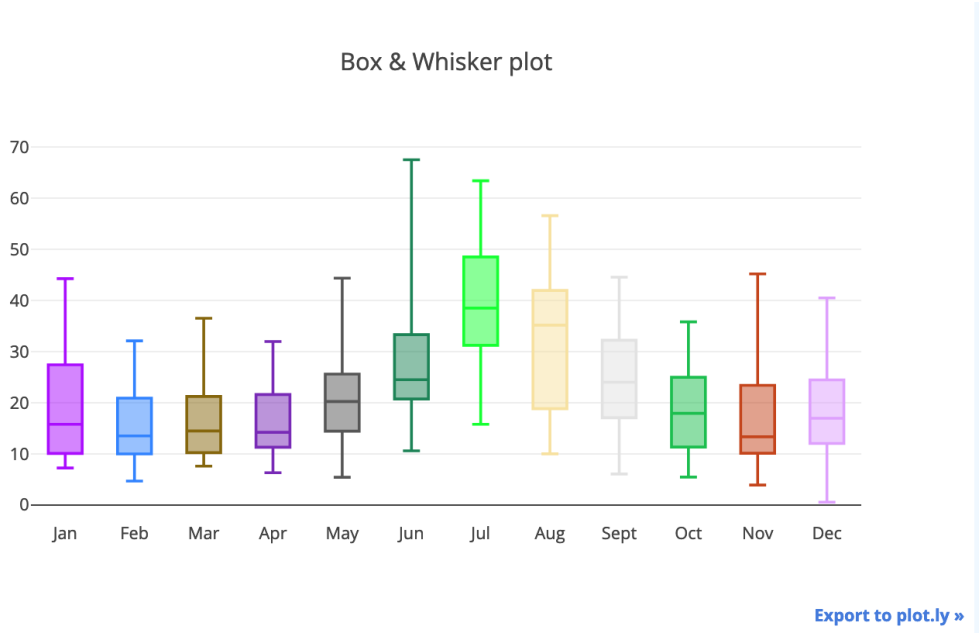

Figure 3. Box and Whisker Plot of Series data.

Three plots gave me some general view about my series data. The observations suggest that the month-to-month fluctuations may not be systematic and hard to model. And the distribution is not Gaussian. Then I decided to manually deploy a configured ARIMA model but the output: ADF Statistic: -7.041553, p-value: 0.000000, Critical Values: 1%: -3.449, 5%: -2.870, 10%: -2.571. The results show that the test statistic value -8.763759 is smaller than the critical value at 5% of -2.872. This suggests that we can reject the null hypothesis with a significance level of less than 5%. Thus I had to do Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots so that I could find out the best pairs for ARIMA parameters p,d,q. below shows the plot:

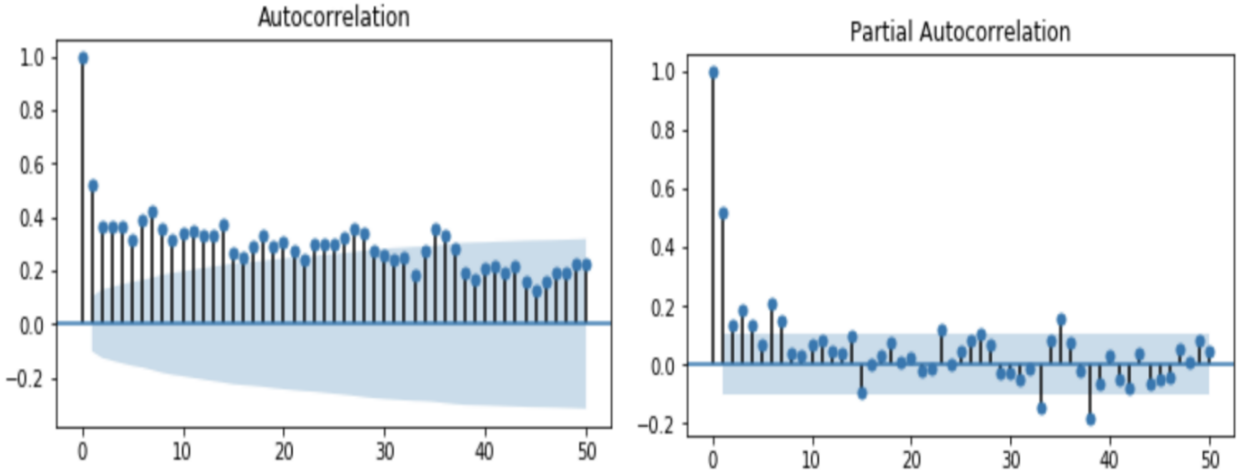

Figure 4. Autocorrelation and Partial Autocorrelation Function plots.

This quick analysis suggests an ARIMA(15,1,2) on the raw data may be a good starting point. The model can be simplified to ARIMA(0,1,2). Then a GridSearchCV had to be implemented for finding the best simplified p,d,q. It is clearly that ARIMA(1,1,1) would be the best would to train the series after search. In order to solve the previous data unbalanced problem, utilized the ARIMA normalization function called Box-Cox Transformation. Then we could get a nearly normally distributed dataset. That is so cool and let us have a look.

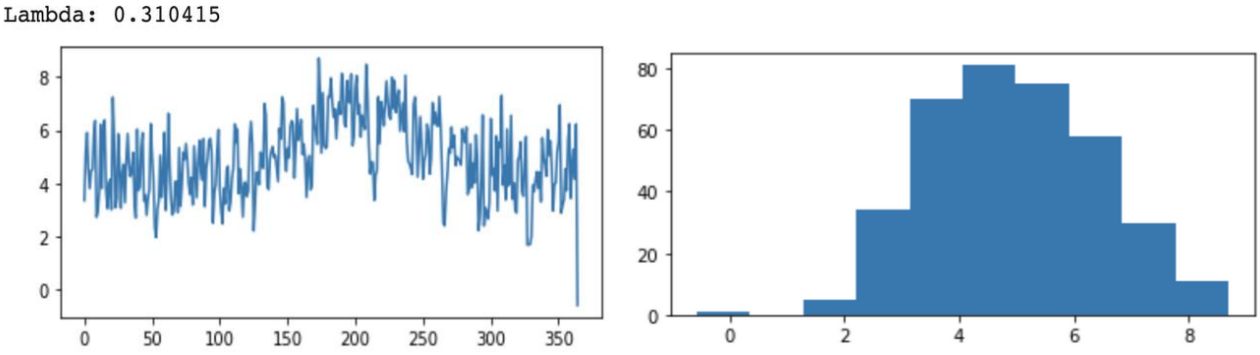

Figure 5. Normally distributed dataset after Box-Cox Transformation.

After doing Box-Cox Transformation, data has shifted from left skewed status to much normal distributed. Then let us print out the best RMSE and the prediction pattern.

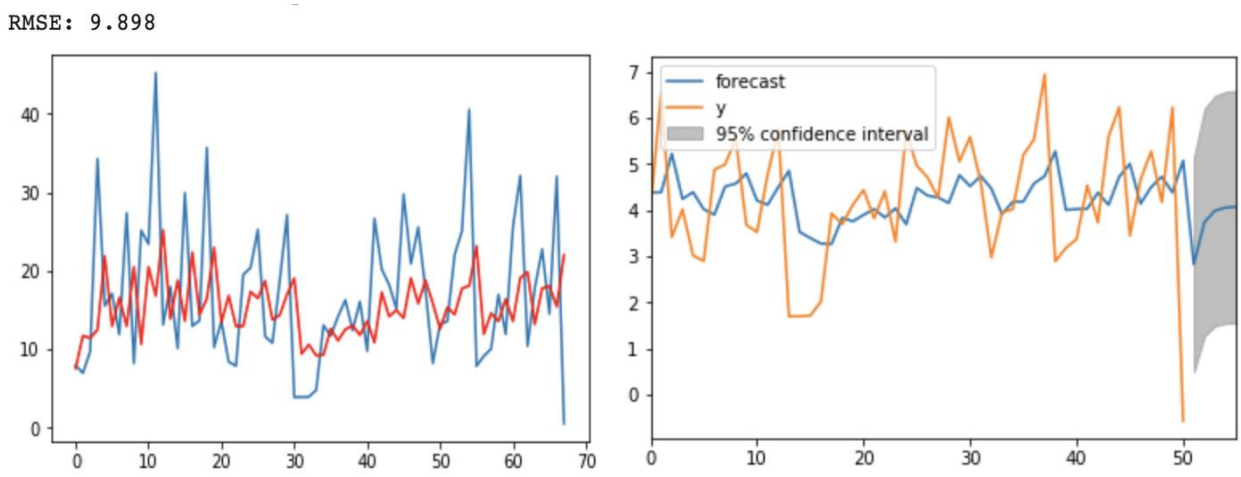

Figure 6. Validation and Prediction.

After all, it is a big improvement that the final RMSE reduced to 9.898 and prediction line is certainly in the 95% confidence interval.

2. Baseline Model

Next, I am going to do EV prediction based on four commonly used machine learning model. Of them four, I basically used the same process: 1. GridSearchCV to find the best_estimator and best_parameter; 2. Print the classification_report, 3. Plot confusion_matrix; 4. Plot roc_auc_score.

Logostic Regression model is more informatic than other classification algorithms. Like any regression approaches, it expresses the relationship between an outcome variable (label) and each of predictors. Aftre doing GridSearchCV, I found the best parameter C is equal to 0.01. The precision, recall, f1-score represents it is generally predict around predict 50% accuracy of label 0 and 1.

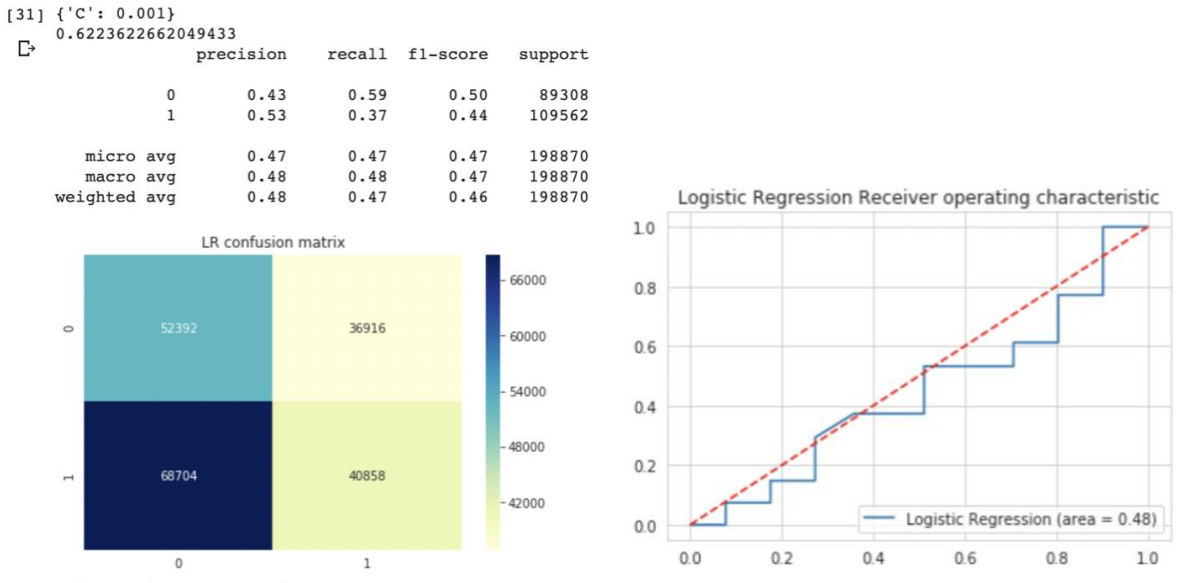

Figure 7. Metrics table, confusion matrix and ROC of logistic regression.

It is not so hard to tell that there are too many type II error which means false negative rate is so high, it indicates that we are supposed to improve the model by using threshold but that depends on the specific situation. The Receiver Operating Characters curve (or ROC curve) is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a diagnostic test. It always demonstrates several things: 1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). 2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. 3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. 4. The slope of the tangent line at a cutpoint gives the

likelihood ratio (LR) for that value of the test. You can check this out on the graph above. 5. The area under the curve is a measure of text accuracy. And it is clear that area of logistic regression is 0.48 by using the best_estimator and best_parameter after GridSearchCV.

Naive Bayes is another technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. I also represented the model by using those plots. Figure 8 provides example pf these plots.

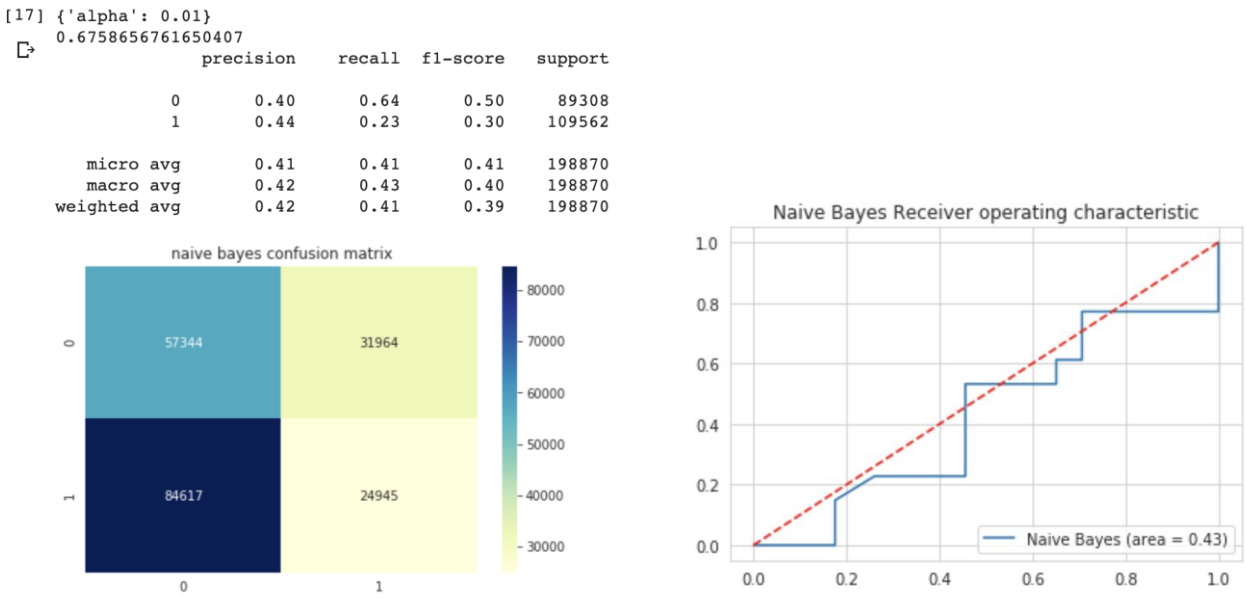

Figure 8. Metrics table, confusion matrix and ROC of Naive Bayes Classifier.

Naive Bayes show a little bit improvement of the label 0 recall which means the true positive is increasing. Though the false negative rate increases also, true positive rates still improves the label 0 recall performance. But overall, the ROC area shows the performance of Naive Bayes is no better than Logistic Regression model.

Random Forest shows the best performance among the four classifiers. And without doubt, it along with the Extreme Gradient Boosting (XGBoost) which I will talk later, demonstrates their high classification ability. Because its hierarchical structure and complexity of the combination, it always shows great result in classification problem. Below is the example performance plots of Random Forest Classifier.

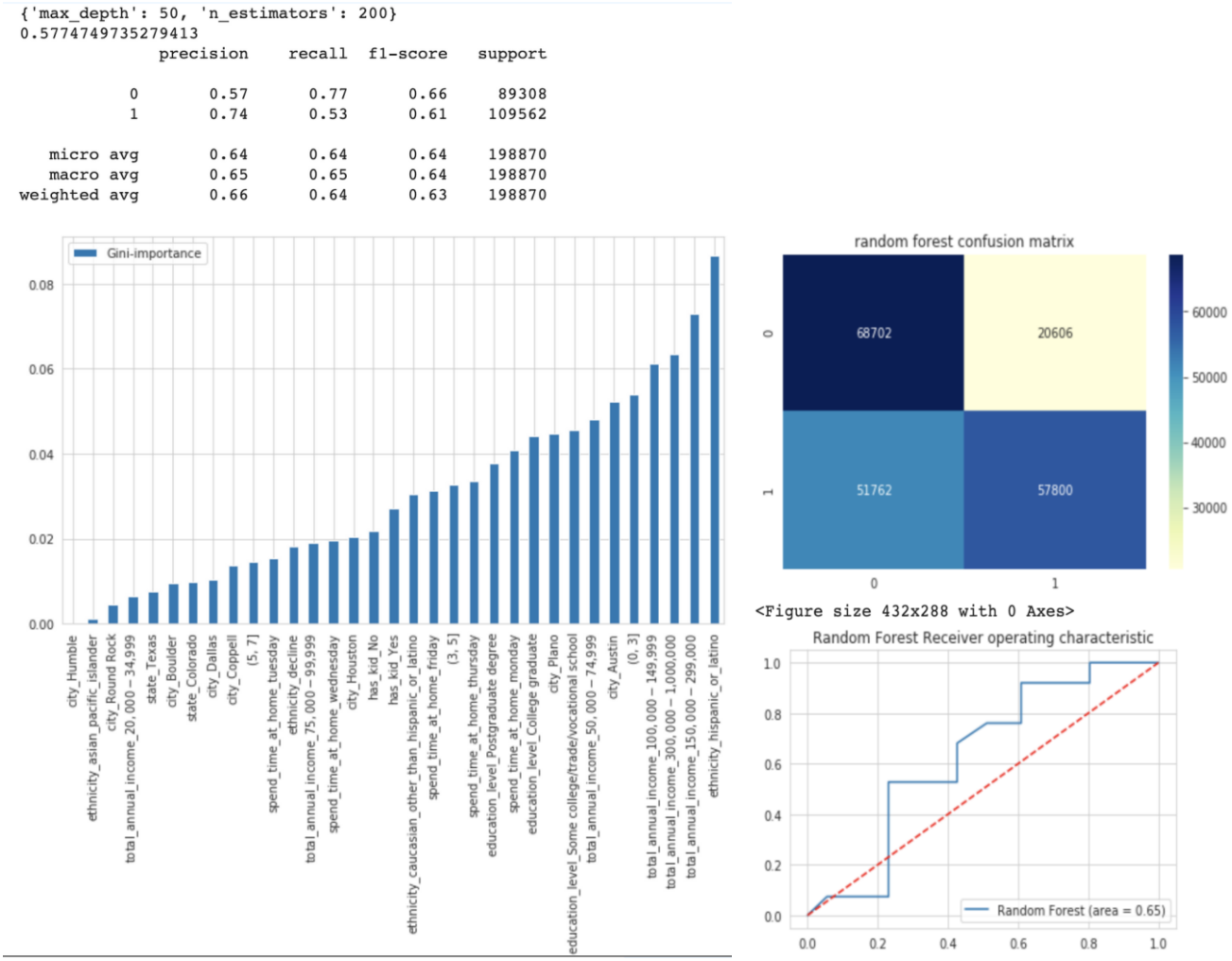

Figure 9. Metrics table, feature importance, confusion matrix and ROC curve of RF.

With more detailed information about my best classifier, I also demonstrate the feature importance in these graphs. And because it is measurement of a tree model, I use Gini importance. This provides some facts that city factor is the least important feature as it meets our common sense that these cities are all in our research area and we could simply treat them as one high similar feature, thus there are no big difference between various cities. It seems that the top 3 important features are 'ETHNICITY_HISPANIC_OT_LATINO', 'TOTAL_ANNUAL_INCOME_150000-299000', 'TOTAL_ANNUAL_INCOME_300000_1000000' which indicates high annual income would be a key factor for purchasing an electric vehicle. It is obvious that true positive rate increases as more False Positives have been correctly predicted as True Positive. This transition is the most advantage of Random Forest Classifier.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

With more parameters to tune such booster (gbtree, gblinear or dart), silence, verbosity, nthread, it took a long time to run out the final GridSearch. So I used default setting to train my own model which gives a equally good performance as Random Forest.

	precision	recall	f1-score	support
0	0.58	0.67	0.62	89308
1	0.69	0.61	0.65	109562
micro avg	0.64	0.64	0.64	198870
macro avg	0.64	0.64	0.64	198870
weighted avg	0.64	0.64	0.64	198870

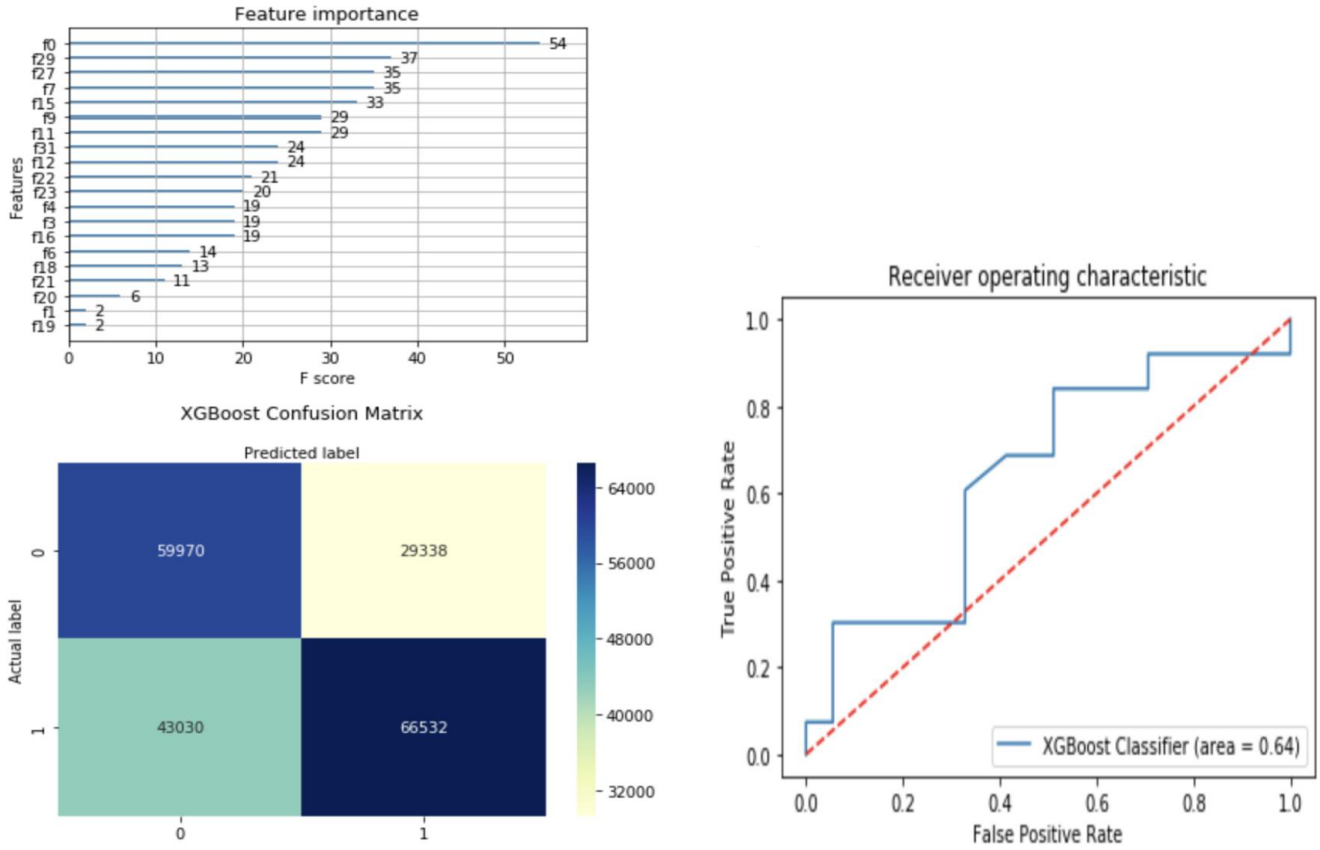

Figure 10. Metrics table, feature importance, confusion matrix and ROC curve of XGBoost Classifier.

Compared with the Random Forest, XGBoost performs much better on distinguishing False Negative as True Positive. Another prove is that label 1 precision is optimized.

RECOMMENDATIONS

The forecasting electric vehicle presence in one household could simply categorized as classification problem by creating training set and testing set that share the common factors of family background information. Then several models solely focused on different advantages have been built to make prediction of unseen families based on family common features. With time series feature such as 'localhour', continuous factor as electric usage per hour could be utilized for building other detection or prediction model. For example, electric usage is predictable with the time factor and EV charging system could be allocated based on the peak hour in different communities or cities. Another distinct approach may be a deeper understanding of household electricity needs, and the development of household smart charging system which leads to a more circular electric system.

CONCLUSION

Electric vehicle presence discovery is the premise of discovering home smart charging network. Through the study of this paper, related industries can take measures to cope with the situation of different family users, such as equipping families who own electric cars with family integrated smart charging system, for families without electric cars, so as to promote green travel and environment protection concept. There is certainly a mix of potential factors which somehow effects the probability of purchasing a electric vehicle such as family annual income, education level, their family size, how many kids do they have, etc. The feature engineering work attempted to make the best factors combination for my model and also avoid the data overfitting and imbalanced for the most part. Toward deployment process, I make the most use of time series model to better handling family electric usage prediction. By analyzing the electric usage, I could predict the future usage pattern so that we could adjust the measurement to some extent. Then a strong linkage line has been built, that is to say, we have multiple solution when dealing with two different kind of families. The accuracy results of the models were low, and therefore, data complexity and number of features after one-hot encoding may lead to this result. Anyway, best model still makes effective classification with both True Positive and True Negative. For future optimizing, one possible solution is to create threshold in order control the False negative rate. And it could also be taken into consideration to reduce the less important ratio to largely reduce the model complexity. By doing so, support vector machine could probably be used for stronger classification.

REFERENCES

Rajashekara, K. (2013). Present Status and Future Trends in Electric Vehicle Propulsion Technologies. *IEEE Journal of Emerging and Selected Topics in Power Electronics,1*(1), 3-10. doi:10.1109/jestpe.2013.2259614

T. Trigg et al., "Global ev outlook: understanding the electric vehicle landscape to 2020", *Int. Energy Agency*, pp. 1-40, 2013.

C. Bikcora et al., "Prediction of availability and charging rate at charging stations for electric vehicles", *Proc. IEEE Int. Conf. on Probabilistic Methods Applied to Power Syst. (PMAPS)*, pp. 1-6, 2016.

L. Gan, U. Topcu, S. H. Low, "Optimal decentralized protocol for electric vehicle charging", *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 940-951, May 2013.

T. Gjosaeter et al., "Security and privacy in the semiah home energy management system", *Proc. EUROMICRO conf. on Digital System Design (DSD)*, pp. 27-29, Aug. 2014.

Energy, 2017, [online] Available: https://www.tesla.com/energy.

Smart community neighbourhoods, 2018, [online] Available: http://smartnaboel.no/.

Electro-erosion and electric-spark machining; Present status and future prospects. (1958). *Wear, 1*(6), 526. doi:10.1016/0043-1648(58)90645-8

Wang, Y., Dai, X., Liu, G., Wu, Y., Li, Y., & Jones, S. (2016). Status and Trend of Power Semiconductor Module Packaging for Electric Vehicles. *Modeling and Simulation for Electric Vehicle Applications*. doi:10.5772/64173

Kikkawa, T. (n.d.). Present status and future trend of low-k dielectrics/interconnect technology for ULSI. *7th International Symposium on Plasma- and Process-Induced Damage*. doi:10.1109/ppid.2002.1042632

Hamada, K. (2011). Present Status and Future Prospects for Electronics in Electric Vehicles/Hybrid Electric Vehicles and Expectations for Wide-Bandgap Semiconductor Devices. *Silicon Carbide*, 1-19. doi:10.1002/9783527629077.ch1

Hamada, K. (2008). Present status and future prospects for electronics in electric vehicles/hybrid electric vehicles and expectations for wide-bandgap semiconductor devices. *Physica Status Solidi (b), 245*(7), 1223-1231. doi:10.1002/pssb.200844079

APPENDIX A

Feature Variable	Data Type	Description
Dataid	Int	Representation of each family
label	Int	1 means electrical household, 0 means non-electrical household
Educational_level	Object	Categorized as College graduation, Postgraduate degree,some college/trade/vocation
Total_annual_income	Int	The annual salary of each family
Ethnicity	Object	Ethnicity of survey participants
Has_kid	Object	Whether they have kids in the family
(0,3) (3,5) (5,7)	Int	Family size
City	Object	City
State	Object	State
Spend_time_at_home	Int	How many days do you spend at home in a week
ssd	Int	Human comfort created by (temperature, humidity, windspeed)