Application of Bayesian Additive Regression Trees to Analyze The Growth of United States Electric Automobile Industry

P Ajidarma1 and D Irianto1

1Industrial Engineering Study Program, Bandung Institute of Technology, Jl. Ganesha No.10, Coblong, Bandung 40132, Indonesia
ajidarmap@itb.ac.id

Abstract. United States electric automobile industry has boomed significantly in the past recent years. Multiple research tried to analyze how a multitude of socio-economical factors correlate with and influence the growth of electric automobile market. This research aims to apply Bayesian Additive Regression Trees (BART) to study the relationship between each factor and the sales of electric vehicle, a proxy of the electric automobile market growth. The full posterior inference feature of BART enables the model to analyze the marginal effects of each predictors. The predictive BART models is selected based on the comparison of the in-sample fit and out-of-sample accuracy. Among sixteen independent variables, the model manages to identify key predictors and important interactions, which are consequential for the decision making in the electric automobile industry.

1. Introduction
Electric Vehicle (EV) sales in the United States are growing at an increased pace with estimatedly 59% growth annually. Factors like zero carbon emissions, lower vehicle operating costs, and incentives given by the government has helped in boosting the growth in electric vehicle sales. Understanding the nature of growth in the EV market provides a pathway to better predict future EV business, and ultimately, the future of automobile industry.

Researchers have been developing sensible prediction models with appropriate factors affecting the growth as predictors to predict sales in Automotive Industry. The modelling approaches often includes Time-series models, Multivariate Analysis, and other various statistical learning models. Now that EV gives the new definition of future automobile industri, this research is focused on to the EV sales prediction, including factors that influence it significantly.

Electrical vehicle is a potential solution for imminent future energy and environmental crisis. The dynamic nature of automobile market, growing volumes of the data, and the supply of computing technology calls for the need of a robust predictive model. This research aims to build a predictive model for Electric Vehicle Sales in US based on a variety of socio-economical factors. The literature reviewed so far offers models that are either good at prediction accuracy or interpretability. But, they could not produce reliable and interpretable forecasts. This study is to fill that gap and build a reliable and interpretable model by using a method named Bayesian Additive Regression Trees (BART), a model that offers robust predictive capability as well as good interpretability.

2. Literature Review
2.1 Predictive Models on Automobile Industry

An analysis by Searle, et al [1] tried to investigate leading factors that affect EV market in California cities. The study found that the comprehensive policy support, local promotional activities and charging infrastructure in California encourage the EV sales growth. A similar study on modelling the EV sales is conducted by Florida Solar Energy Center, as published in Electrical Vehicle Transportation Center Report [2]. The model used EV annual sales as response variable. It also harnessed the cumulative quantity of vehicles on the road between 2010 and 2014 and the growth rates during these 5 years. Using these data, the sales prediction for the next 10 years is made.

Multiple Linear Regression (MLR) is used to forecast vehicle sales, as observed in a study by Shahabuddin [3]. The study examined the relation between automobile sales and a variety of economic and demographic variables, such as Gross Domestic Product (GDP), Gross National Product (GNP), population, personal consumption, industrial demand, and institutional funds, both short-term and long-term. There is strong relationship between foreign car sales and economic variables. Nevertheless, such relationship is nonexistent for domestic car sales, proven when he examined the incentives used by the domestic car makers to influence car sales.

Other study tried to tackle this problem using a combination of statistical learning methods. Bru’hl, et al [4] combined basic time series methods with a trend estimation calculated by Multivariate Linear Regression (MLR) or a Support Vector Machine (SVM), with a Gaussian kernel to predict the registrations of new vehicles in the German automobile market. Ten market influencing factors are included in the model, which describes the global/national economy, the characteristics of the automobile market, the consumer behavior with respect to changing economic cycle and the influences of credit restrictions. Furthermore, Hülsmann M, et al [5] took the same data set and added two additional indicators for US-specific market, which are Dow Jones and BCI. The study used a variety of methods: Ordinary Least Squares (OLS), Quantile Regression (QR), SVM, Decision Trees, k-Nearest Neighbor (KNN), and Random Forest (RF) to make the prediction.

Further development of predictive modeling includes optimization of the input parameters. For instance, Particle Swarm Optimization (PSO) has been used to determine the inputs for Support Vector Regression (SVR) in Lu, et al [6]. The required three input parameters of SVR are tube-width in the loss function (ε), standard deviation of the kernel function (σ), and cost coefficient for experimental and structural risk evaluation (C). It has also been concluded that PSO-added SVR performed better in terms of prediction accuracy compared to similar simulation models, such as Genetic Algorithm with SVR.

2.2 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) is initially developed by Chipman, et al [7]. BART is a bayesian-based nonparametric regression which harnesses dimensionally adaptive random basis elements. Based on ensemble methods and boosting algorithms, BART is defined by a statistical model consisting of a prior and a likelihood. BART enables full posterior inference including point and interval estimates of the nonparametric regression function, including the marginal effects of the predictors. By keeping track of predictor inclusion frequencies, BART can also be used for model-free variable selection. We could state the supervised learning problem as,

\[ Y = f(x) + \varepsilon \]  

where \( \varepsilon \) is the error term and follows normal distribution with mean zero and variance \( \sigma^2 \). The main goal is to estimate the f(X). In BART,

\[ Y = f(x) + \varepsilon = g(x, T, M) + \varepsilon = \sum g(x, T_j, M_j)_{m_j} = 1 + \varepsilon \]  

where T denote a binary tree, and M=\{\mu_1, \mu_2, ..., \mu_b\} denote a set of parameters associated with each of the b terminal nodes of T. The \( (x,M_j) \) represents a regression tree, and m is the total number of trees in the model. The main idea of BART is weakening the \( (x,M_j) \) effects by imposing a prior
that regularized the fit by keeping the individual tree effects small. Then, BART ends up with a sum of trees, each of them explains only part of the variation of \( f \).

The fitting and inference are accomplished through an iterative Bayesian back-fitting MCMC algorithm that generates samples from posterior distribution [9]. BART is a full Bayesian specification, the information of all unknown parameters \( \theta=\{T_1,M_1,\ldots,(T_m,M_m),\sigma\} \) would be estimated by the posterior distribution:

\[
\pi(\theta|Y) \propto p(Y|\theta)\pi(\theta)
\]

Bayesian methods are popular in many areas, because they are robust to over-fitting in the presence of small sample sized and can handle missing or incomplete data. BART also have those benefits. BART enables a full assessment of prediction uncertainty while remaining highly competitive in terms of prediction accuracy.

3. Methodology

This research aims to use BART to study how the Electric Vehicle (EV) market is influenced by several predictors stated by Ajidarma, et al [8]. The predictors are as follows:

| Variables | Description |
|-----------|-------------|
| gini      | Index of income inequality |
| income    | US median family income in |
| travel    | Mean travel time to work |
| charge    | Number of charging infrastructures |
| edu       | Percentage of bachelor’s degree holder |
| gas       | Price of a barrel of crude oil |
| Pubtrans  | Availability of public transportation |
| Air       | Amount of CO2 in the air |

| Variables | Description |
|-----------|-------------|
| resexp    | Amount of grants given for research |
| employ    | Proportion of unemployment |
| vehcap    | Number of privately-owned vehicles |
| elcost    | Electricity cost |
| noevmnt   | Number of EV models offered to each state |
| incent    | Number of variety of incentives given in each state |
| popdens   | Population density for each state |
To fit the sum-of-trees model, BART uses a tailored version of Bayesian backfitting MCMC algorithm [9] that iteratively constructs and fits successive residuals. BART differs with gradient boosting approach [10] in how it weakens the individual trees and how it performs the iterative fitting. BART algorithm undergoes a burn-in period, and afterwards the algorithm converges and produces a sequence of $f^*$ that can be deemed as dependent sample of $p(f | y)$. This research develops four BART models based on differing parameters which are designed as follows:

| Parameter               | Description                                                                 | Value          |
|-------------------------|------------------------------------------------------------------------------|----------------|
| Number of Trees         | Number of trees to be used in iterative backfitting algorithm                 | 150            |
| $q$                     | Quantile of the prior on the error variance at which the estimation is placed; larger $q$ gives more aggressive fit | 0.75 and 0.99  |
| Number of Burn-in Samples | Number of MCMC samples to be discarded during the burn-in period             | 10 and 100     |
| $k$                     | The prior probability that $E(Y | X)$ is contained in the interval $(y_{min}, y_{max})$, based on a normal distribution | 0.95           |

Each model is tested using four indicators that measure the goodness of fit. R-Square and Train RMSE measure how well the prediction is made on the training set, while Test RMSE measures the goodness of fit between the predicted values and the actual data of the testing set. Further, Cross Validation Error (CV Error) measures how the results of a statistical analysis will generalize to an independent data set. In other words, a model with low CV Error is less prone to over-fitting towards the train data.

The experimental results, as shown on Table 2, shows the performance of different BART models with its corresponding tuned parameters. On observing the performance characteristics, the fourth BART model (BART4) seems to be the best model compared to the rest of the models, based on all four indicators. Thus, further analysis of the influencing factors is conducted based on BART4 model.

| Parameter     | Description                                                                 | Value          |
|---------------|------------------------------------------------------------------------------|----------------|
| Rsq           | TrainRMSE                                                                   | 0.9965         |
| TrainRMSE     | TestRMSE                                                                    | 0.000008       |
| TestRMSE      | CVError                                                                      | 0.000060       |
| CVError       |                                                                              | 0.000191       |

4. Analysis
Diagnostics on the error variances in the Figure 1 show that residuals are normally distributed and there are no patterns observed in the residual plot. There are notable a few outlier points, which are the data from states with exponentially larger sales, such as California. However, the majority of the points seem to be densely distributed at one end of the plot.
The variable importance plot of BART4 model is shown in Figure 2. The top predictors are noevmod, charge, pubtrans, elcost, travel, edu, popdens, incent, and gini. This research also analyzes the interactions between different predictors in the model, shown in Figure 3. It shows that the interaction between pubtrans and incent has the highest significance, followed by resexp and charge. The interactions are intuitive for the inherent dependencies of each other in the real world.
5. Conclusion
This research is conducted to gain further understanding regarding inference and prediction of EV market for each state in the United States. Fifteen predictors are initially proposed into each of our model fitting: gini coefficient index, family income, mean travelling time, number of charging stations, education rate, crude oil price, availability of public transportation, air quality, research expenditure of each state, unemployment ratio, vehicles per capita, cost of electricity, number of EV models offered in each state, governmental incentives, and population density. The chosen dependent variable is annual per capita sales of Electric Vehicle in the United States.

Four Bayesian Additive Regression Trees (BART) models are generated and fitted into the data. Furthermore, the fitting quality is measured based on 4 indicators: Training RSME, R-Square, Testing RMSE, and Cross-Validation (CV) Error. Based on the optimal model, important variables are concluded to be the number of EV models offered in each state, number of charging stations, availability of public transportation, cost of electricity, and mean travelling time. These factors are proven to correlate significantly with US EV sales, and further study should try to answer how they will be influential towards the future growth of the electric automobile industry. Lastly, interactions between predictors are also identified in the analysis phase.

References
[1] Searle S, Pavlenko N, and Lutsey N 2016 Int. Council on Clean Transportation.
[2] Bru’hl B, Hu’lsmann M, Borscheid D, Friedrich C, and Reith D 2009 ICDM. Lncs 5633 146–160.
[3] Syed S 2009 Management Research News 7 670 – 682
[4] Hülsmann M, Borscheid D, Friedrich C, and Reith D 2012 Trans Machine learning and Data Mining 2 65 – 86
[5] Lu X and Geng X 2011 Int. Conf. on Intelligent Computation Technology and Automation 1 71-4.
[6] Block D, Harrison J and Brooker P 2015 Electrical Vehicle Transportation Center Report.
[7] Chipman H A, George E, and McCulloch R 2012 Annals of Applied Statistics 1 266-298.
[8] Ajidarma P, Gong S, Nanduri L, and Nateghi RA. Proceedings of the Transportation Research Board 97th Annual Meeting.
[9] Hastie T, Tibshirani R, and Friedman J 2009 The elements of statistical learning: data mining, inference, and prediction (New York: Springer-Verlag).
[10] Friedman J and Greedy 2001 The Annals of Statistics 29 1189–1232.