Executive Summary

This research project focuses on the classification problem that may enhance a utility companies' knowledge of its consumer's needs. The aim of this research is to build a machine learning algorithm that can discover the presence of an Electric Vehicle. The algorithm will be feed features that describe homes, such as energy use, age of house and size of home to predict if there is an Electric Vehicle. This project used multiple algorithms to find the best model. These algorithms are Decision Trees, Neural Networks, Random Forest, Logistic Regression, K-Nearest Neighbor and Support Vector Machine. This work may bring more insights into the consumers of utilities companies. However, these data did not provide enough evidence of a strong relationship to the response variable. Upon further investigation, after no amount of feature engineering and hyperparameter tuning brought the accuracy to a higher level than simply predicting that no one had an EV, the survey provided by the participants in the electricity data program showed that many homes with EV did not charge their electric vehicles at their house. This explains why using electricity as a predictor of EV presence will not lead to an accurate model.

Preface

As part of Data Science Capstone (DATS 6501), this work describes the data science work performed during the spring semester of 2019. This project was first proposed by Oracle to George Washington University Data Science students. This project was chosen as it pertains to an interest in the environment and environmentally friendly decisions. The task at hand was to create a model that could predict the presence of an electric vehicle. This project was completed using Python Jupyter Notebook. It required streaming data, data preprocessing, statistical analysis, machine learning algorithms and data visualization.

Contents

Introduction

As a utility company, knowledge of the consumers is essential. Understanding how much energy they will require is necessary to have the correct infrastructure to support their needs. Electric vehicles are something that currently many Americans do not own but is marketed to reduce Petroleum consumption. This makes it interesting to certain consumers. As electric vehicles find their way into residential homes, those homeowners or renters will have to charge the vehicles. If residential homes are charging electric vehicles at home than the energy consumption can increase.

This is information that can benefit utility companies. The demand of electricity due to electric vehicles can be uncertain. There has been prior research done on electric vehicle charging demand. According to Bae and Kwasinki (2012) the demand for electric vehicle charging changes in both time and space. The ability to predict when and where consumers will want to charge electric vehicles affects the power grid and is useful in planning distribution of power to the grid. Those that own electric vehicles may charge their vehicles at different times of the day depending on different aspects of their lives. The location may also change as there are various public charging stations, many owners can choose to charge their vehicles at their own homes, and some may have access to an electric vehicle charging station at their place of work. Knowledge gained about the need for electric vehicle power demand can let suppliers avoid power loss and provide overall improvement to the voltage in the smart grids. Utilities companies have concerns over excess stress on the system and how to manage the demand. The excess demand could lead to power losses, overloads and other issues that would lead to consumers feeling that the company is less reliable (Deilami, Masoum, Moses & Masoum, 2011). Therefore, knowledge gained about the whereabouts of electric vehicles can improve the utilities companies planning for the grids.

Narrative

Definition of Problem

The objective of this study would be to recognize which residential homes own electric vehicles. This information would provide more insights into the needs of energy in homes. Electric vehicles require much energy to charge. Being able to predict where electric vehicles are can lead to energy companies have a better estimate of how much energy certain homes need. This can also lead to energy companies being able to provide better energy saving/money saving tips for their consumers.

Research Question

Which residential homes have Electric Vehicles?

Project-Task

The work assignment was to generate machine learning algorithms (classification models) to classify if the home has an Electric Vehicle.

Methodology

This project followed the Data Science lifecycle. It began with business understanding. This entails information regarding energy companies and electric vehicles. The next steps involved gathering data and scraping it for the necessary information. The next step is to clean the data. This data was messy. It came as hourly energy use for each house. This project used three years of energy data. The data was found on Data Port's Pecan Street. The metadata provided a minimum and maximum timestamp for each house in the program. All houses that joined prior to January 1st, 2016 and left after December 31st, 2018 were included in the data stream. This way all houses in the data used for the project were in the program for the entire time that was considered. The energy feature was grouped by the residential home to get a total amount of energy used by each home. This data also came in multiple tables. These tables were merged on the unique identifier, named DataID for each of the residential homes. Residential homes that had mostly null values for the features of interest were dropped. The other features included in the data other than energy use and the presence of EV were chosen to control for variability. For example, the size of the house, the type of building structure, the year of construction and its location may affect how much energy the house uses. Solar panels were also included as there is a strong relationship between people who purchase electric vehicle and people whose homes have photovoltaic panels.

Exploratory data analysis is the next step in the data science lifecycle. This analysis showed how electric vehicle homes differ from those without. The exploratory data analysis leads to feature engineering. The original features were augmented slightly, and multiple data frames were created with the different feature engineered types. Then the different data frames were modeled on random forest and logistic regression. This allowed for an understanding of if the feature was important to predicting the presence of EV. The best of the feature engineered data frames was used for predictive modeling. Multiple algorithms were used on a 5- fold validation to identify the best model. The last stage of the data science lifecycle is data visualization. This is essential for communication and presenting the data science process and results.

The evaluation of the model did cause for optimize to improve by going back to the model building and feature engineering pieces of the data science lifecycle. This did not improve effective in creating better results. This led to an adjustment in data extraction and then more data preprocessing to understand the surveys that participants filled out to provide more insights into those that purchase electric vehicles and their charging behavior.

Data Processing

Exploratory Data Analysis

The main purpose of this part of the analysis is to understand the data. For this project that means how do EV homes and non EV homes differ. Overall there are more homes without EV. This can be seen in figure 1. Figure 1 also shows that the distribution of homes with and without EV do not drastically differ. However, there is a slight difference in that EV homes tend to be bigger, use more energy and are newer.

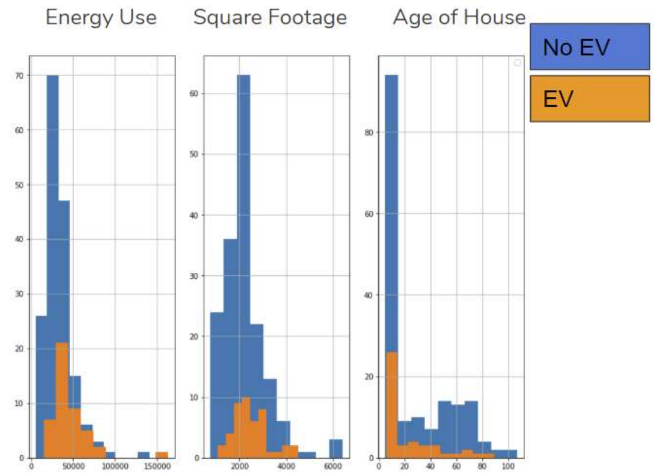

Fig 1. Histogram of continuous variables

There were some data limitations when it came to the categorical variables. For example, for building type and location. There were many options, yet most homes fell under one condition and there were only single digits in the others. Instead of creating a multitude of dummy variables. These variables were turned into the most popular versus other. This made it much easier to understand the features. The odds ratios for each of these features was conducted with the presence of EV. The odds of having an electric vehicle among those with single family home are .57 times the odds of having an electric vehicle among those with our housing types. The odds of having an electric vehicle among those with PV are 5.93 times the odds of having an electric vehicle among those without PV. The odds of having an electric vehicle among those that live in Austin are 2.3 times the odds of having an electric vehicle among those that live elsewhere.

Data Modeling

Feature engineering was a large part of this process. There was the question of how to look at energy use. This was answered by using t-tests, logistic regressions, and feature selection from random forest. The different ways that energy use was dealt with was by totaling all energy use of three years for all homes and comparing EV with non EV to make sure that they were statistically significant. It was. It was even statistically significant at the average hourly energy use level. Figure 2 shows the QQ plot for average hourly energy use. The plots did not appear to be normally distributed on the raw data, and to use a T-test it must be normal to satisfy assumptions. The outliers were removed and the QQ plots do show a relatively normal distribution for energy use. The t-test on average hourly energy use comparing the mean energy use for homes with and without EV returned a T-value of 2.79 and a p-value of .007.

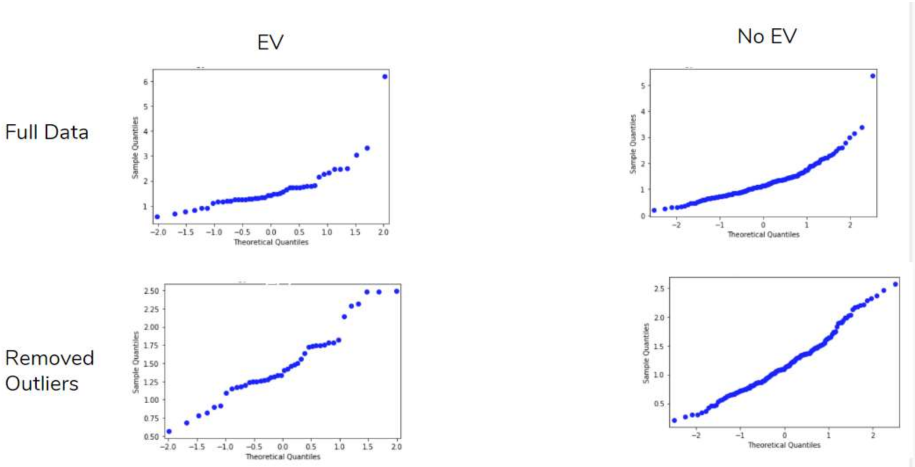

Fig2. QQ plots for average hourly energy use with and without EV

However, when total energy use was divided by square footage of the house it was no longer statistically significant. This was very interested and showed that EV owners having larger homes is a factor into the amount of energy that they use and an explanation for why they use more energy than non EV owners.

There was the question of how to look at year house was built. It seemed more straightforward to convert this to age of home. People tend to think of the age of a house as part of its integrity. It was easier to see that as the number goes up it is older and as it goes down it is newer. When the year is in place, it is on a very different scale and some audiences may feel odd thinking about a lower number as an older home. Converting year house was built to age of house was a way to avoid any potential confusion to potential audience members.

Multiple data frames were created to tryout different versions of these feature engineering changes. Each of these data frames was run through both a logistic regression and a random forest model with a test train split of 30, 70 and on oversampling technique that created equal ratios instead of having a class imbalance.

The accuracy score from the test and train data on these models was used to evaluate the best data frame. Another measurement used was the R Squared value of the logistic regression model, which coefficients were statistically different from 0 based on the logistic regression output and the feature importance that random forest assigned to each feature. This was used as a holistic approach to evaluate the data frames. Comparing each data frame on each of these metrics allowed to multiple ways to provide more confidence that it is the best data frame to move forward with for model building. Figure 3 below shows the feature importance from the random forest output on the data frame that is the best version, which includes total energy for the three years, age of house and has outliers removed. Here energy use is the most important feature.

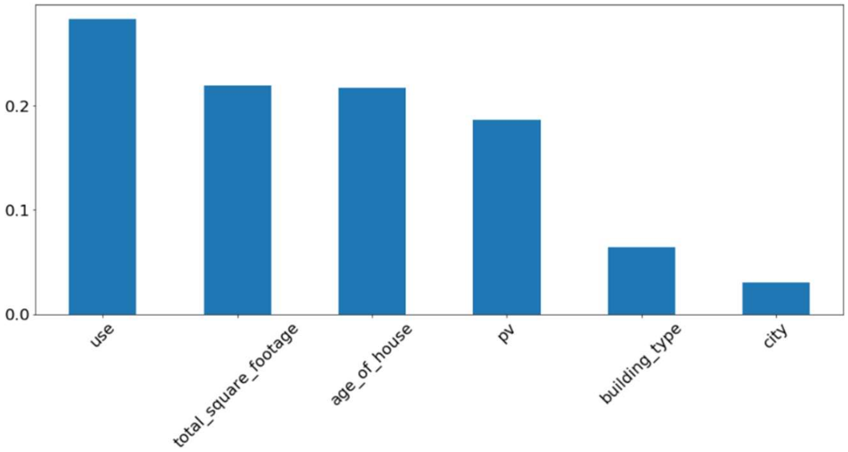

Figure 3. Feature Importance from Random Forest Output

Once the features were selected the problem of class imbalance had to be dealt with. The ratio of EV to non EV in the original data is 20,80. This was used with logistic regression and random forest where the overall accuracy and r squared values were evaluated. Other ratios of 30,70 and 40,60 and 50,50 were also used. This way multiple trials could be done to identify how the model reacted and what happened. These models used a train test split of 70,30. Viewing accuracy of both train and test and the overall significance of the model and viewing confusion matrices it was decided to use 50,50 as the ratio of EV to non EV.

The final step was to create a pipeline that would standardize the data and run it through multiple classifiers under multiple parameters with a change in the features and use a cross fold validation. The features were selected based on the feature importance from random forest. The threshold was put in place to have each selection of parameters for each classifier to run the model with all features, the top three features, the top two features and then just the best feature. For each classifier hyperparameter tuning was utilized through a parameter grid search. The classifiers used were Decision Tree, Random Forest, Logistic Regression, SVM, Neural Network and KNN. The cross-fold validation was do with 5 folds.

Results and Evaluation

Originally oversampling had been done as part of the data processing. Therefore, the results from the models had bias. The first sign of bias was that the test set had higher accuracy than the train. This is unusual. To test for bias the models were run again but with oversampling only on the train set. This way the test set is just the original raw data. This showed a significant decrease in the accuracy on the test set. Another method of working to remove bias was to do under sampling on the train set. The under sampling showed poor accuracy on both the train and test set. This confirmed that oversampling the data does put a bias into the data. The model that showed great accuracy was flawed.

One thought on how to improve the model was to put the energy data in by day instead of the total energy use of three years. There is variability among how much energy is used over time, by combining that data into one feature the model became simpler but much data was lost. To test this theory, one month of energy data was run through the algorithms. To add additional data that could help to explain the variability in energy use among homes was the feature that displays the number of appliances and outlets that the home had. This new data frame was performed on a train test split with under sampling. This model also had issues. It overfit on the train and immediately ran back with a 100% accuracy. On the test data is simply predicted that no one had an electric vehicle. This showed that adding in those additional features did not provide a better model.

Even though the original results did at first glance appear to be a good model, under further evaluation it was a biased model. Multiple other models were run changing a few characteristics, however none of these models panned out well. This led to the question of why these models are not great. To answer this a deep dive into the survey data was conducted. The surveys are not completed by many people. It appears that just because a person opts into the pecan street program, does not mean that they will fill out the survey. The survey is realistically just a sample of the residents in the program. The survey has many questions about the habits and opinions of the members of the program. This includes what temperature do you keep your home at and do you support developing renewable energy and do you have friends living with you. The analysis of these answers that compared the homes with EV and the homes without did not show any strong pattern or difference in those homes. The question that did answer the question of why the model did not pan out, is Do you charge your electric vehicle at home? The majority of EV owners said no; they never charge their electric vehicle at home. Thirteen people said always, which is 27% of all people who answered this question. Therefore, the model was trying to predict who has an electric vehicle based on electricity used, which is not a good predictor if most people do not charge it at home. This question explains why energy use will not be able to predict who has an electric vehicle.

Conclusion

The biggest takeaway from this project was to check the underlying assumption prior to starting the project. In the initial meeting where the project was discussed, the client seemed very confident that everyone with electric vehicles charged them at home and that they took up plenty of extra energy. This was a flawed assumption and is the reason that the model does not predict accurately. For future project, establishing all prior assumptions and verifying their accuracy is a first step. Understanding the accuracy of all assumptions can severely sway the methods that are taken to deal with the problem at hand.

There are many reasons that someone would purchase an electric vehicle. It is entirely possible that having easy access to public charging stations or having charging stations available at work could be a contributing factor into a consumer deciding to purchase an electric vehicle. There are many different reasons why electric vehicles are purchased and many potential benefits. One such benefit is the reduction of petroleum and CO2 emissions (Himelic & Kreith, 2011). This suggests that environmental awareness may be a factor in electric vehicle purchase. This could mean that those who purchase electric vehicles act with environmentally responsible and take steps to reduce their overall household electricity usage. There is the possibility that taking the extra steps to reduce energy use is offsetting some of the extra electricity needed to charge the vehicle. The initial assumption that the presence of an electric vehicle would cause a large spike in energy use and be related so strongly to EV presence that it would be the only feature needed to predict presence of EV was not correct.

The other takeaways include a gain in understanding of the data science lifecycle. This project was on course and at the end of the model building, a return to the database was needed. A full understanding of how data can be manipulated and conducted was a learning. This project also showed that parameters and algorithms are not enough to build a decent model. The data that goes in is the most important factor and if those features are not good predictors, a good model will not be created.

Lastly, a learning on bias and overfitting occurred. The oversampling method to deal with class imbalance lead to a bias in the data and this created a model that overfit. The model used a 5 cross fold validation, but the oversampling biased the model more than the cross validation worked on splitting and learning from the data. The takeaway here is that more data is better. It would be best if there were enough points to do under sampling than oversampling. This would avoid creating a bias within the data.

Future Research

To continue understanding homes with electric vehicles. It would be beneficial to find a data source that contains home energy use and other features (demographics of the individuals as well as information regarding the physical structure) before and after the electric vehicle was purchased. This could provide better insights into how an Electric Vehicle affects the energy use of the home. Ideally, the only changing factor would be the addition of an Electric Vehicle and all else would remain equal (i.e. the number of people living in the home and their energy consumption habits, no remodeling or changes to the physical structure of the home). This data would provide additional resources to the exact nature of how energy needs change with the presence of an electric vehicle. Of course, this data would need to be on homes where the vehicle owner charges their vehicle at home most of the time.

Another way to continue the research would be to investigate the types of batteries and charging stations used by the electric vehicle owners. The different battery types, and charging stations having differing energy needs. This additional information would provide insights into the energy use of electric vehicles and if utility companies know who has this car, and how they charge it. Then they can predict how much energy they will need with more precision.

If possible getting data on more homes would be helpful. Overall, this dataset only contained 13 residents that charge their vehicle at home on a regular basis. If this number was larger than a classification problem could be completed. However, this classification would be not for who owns EV, it would be for who charges EV at their home on a regular basis.

Value of Experience

This was a very interesting opportunity as a student. It was a learning experience. Being able to work with real world data as a student is an invaluable opportunity. It came with many challenges and hardships that proved difficult to overcome. Real world data is incredibly messy, much messier than any other class assignment. Every step of the data preprocessing was a thought exercise. Every line of code was a decision. This assignment solidified that data science has many paths and each one had positives and negatives. Every decision when working with this data was weighted on how making this change can push it forward, but also how it could cause limitations of the data. This data was especially difficult due to its small number of houses contained in the database.

Ideally, in any classification problem the more data the better. For this data there were many null values. Deciding to drop a home because it had missing data was difficult because it seems unwise to lower the total number of observations. In the end a few homes were dropped because they were missing too much information and the goal of this project was not to generate data but to classify it. This data has other issues, such as that it was an imbalanced dataset. Many different oversampling ratios were used to try to assess which would lead to the best model. In this case 50,50 ratio seemed the best because it led to the most cases of predicting yes on EV.

Overall, this data provided many struggles. There were so many data issues and limitations that it turned from “identify problem and find solution” to think of anything that could cause this problem and think of anything that could be a solution and try everything. This was a time management exercise as well as a data science project.

This project was a learning experience. It required knowledge from all previous data science courses taken and gave a practical application of those classes. It was an interesting experience.

References

S. Bae and A. Kwasinski, "Spatial and Temporal Model of Electric Vehicle Charging Demand," in *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 394-403, March 2012.
doi: 10.1109/TSG.2011.2159278

S. Deilami, A. S. Masoum, P. S. Moses and M. A. S. Masoum, "Real-Time Coordination of Plug-In Electric Vehicle Charging in Smart Grids to Minimize Power Losses and Improve Voltage Profile," in *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 456-467, Sept. 2011.
doi: 10.1109/TSG.2011.2159816

Himelic JB, Kreith F. Potential Benefits of Plug-In Hybrid Electric Vehicles for Consumers and Electric Power Utilities. ASME. *J. Energy Resour. Technol.* 2011;133(3):031001-031001-6. doi:10.1115/1.4004151.

Data Source: Pecan Street Inc.