Final Project Report

Electric Vehicles Charging Stations

Predictive Analytics

DHIMAAN SARKAAR
Principal Strategist
& Data Scientist

Executive Summary

For this project, we decided to learn more about the rising trends in the purchase of the Electric Vehicles (EV) and its necessary infrastructural requirements. Hence, we decided to study about the fundamental infrastructure needed for the smooth operation of electric cars. As a result, we directed our focus on researching and predicting the number of EV charging stations required in each county. Over the course of the study, it was our primary focus to analyze and derive insights from the master dataset we created. Our dataset was consolidated from 61 different datasets into one master file. The dataset includes column variables including across various demographic, geographic and socio-economic factors. In order to initiate the process, we first began by began by data engineering where data gathering, consolidation, cleaning and transformation took place. Once we had our final dataset ready, we underwent exploratory analysis using various data visualization techniques. Next, we each picked individual machine learning models, where the important features were selected and different methodologies were explored. The three main models used for this study are Linear Regression, Decision Treen K-Nearest Neighbor and Recommender System. All the three machine learning models were then compared and the best performing model was chosen. In this case the best model was of the Decision Tree as it best satisfied the objectives of project.

It would thus be interesting to learn about the number of charging stations that each county in the United States require based on the current market trends. The initial findings from the dataset showing that our KNN model, accurately predicts the number of EV stations based on the feature similarity. It is a great model for comparing the selected features, allowing us to decipher additional information. Lastly, we were not only able to derive crucial insights, but also provided effective solutions to the business for improved business performance. It would not only add value to the EV industry but the society at large due to improved environmental decisions.

Table of Contents

Introduction

About Project

In our recent years, we have seen a sudden shift in the public perception and upcoming trends. One such example of a rising trend is the increase in the purchases of Hybrid and Electric Vehicles. This new product category has certainly drawn a lot of consumers and has created an increased demand for such automotive goods. Hence, our project is focused on studying the future of this growing category and forecast the necessary infrastructure such as EV Charging Stations.

As a background study, we conducted a literature review on “The Charging Behavioral Analysis” along with “Predicting its Future Demand”, published by ‘The Institute of Electrical and Electronics Engineers ’also known as IEEE. The study was largely inconclusive due to the lack of high-dimensional dataset and less data points. Hence, understanding the limitations, we created our own dataset across various primary and secondary sources. This would ensure that our final dataset has enough datapoints, and contains the necessary factors to predict our desired outcomes.

Literature Survey

In our recent years, we have seen a sudden shift in the public perception and upcoming trends. One such example of a rising trend is the increase in the purchase of Hybrid and Electric Vehicles. This new product category has certainly drawn a lot of consumers and has created an increased demand for such automotive goods. Hence, our project is focused on studying the future of this growing category and forecast the necessary infrastructure (EV Charging Stations) required to meet the predicted demand. Hence, a good strategic plan in synergy with having a thorough understanding of the future trends in the automobile industry is of the paramount importance. Increasingly we have noticed that a lot of the automobile manufacturers are introducing more electric/hybrid vehicles and reducing their reliance on fuel combustion technology. The international focus on Climate Change has forced a lot of companies to adopt a more eco-friendly business model and explore a business avenue that is more environmentally

As per the research study conducted by the “IEEE” in September 2020, It was found that the Machine Learning models along with Deep Neural Networks, were used to carry out the charging behavior analysis and predicting its future demand . Although the study was not able to suffice its desired quality of research due to lack of high dimensional and EV charging datasets. Hence the ML models lacked sufficient training due to insufficient data (Shahriar et al., *Machine Learning Approaches for EV Charging Behavior: A Review* 2020). At the time of the study there were only two publicly available datasets and the rest were all held by commercial companies. In order to have a better study for a research project it would be important to find the latest possible datasets which would consist of a large set of records along with being high-dimensional. We would also have to use more than one dataset so consolidate all of the required data for an accurate predictive model.

In addition, the global market share of Electric Vehicles is predicted to dynamically grow over the course of the next decade and make the industrial-age vehicles obsolete. According to a detailed study conducted by Deloitte, “Global EV forecast is for a compound annual growth rate of 29 per cent achieved over the next ten years: Total EV sales growing from 2.5 million in 2020 to 11.2 million in 2025, then reaching 31.1 million by 2030” (Woodward et al., *Electric vehicles Setting a course for 2030* 2020). Hence, it is imperative for us to accurately predict and plan for the future trajectory. Therefore, an in-depth study is essential to plan for the infrastructure necessary to induct this newer green technology. The study by Deloitte Insights suggests that not only would the EV be instrumental in reducing fuel emissions but also make the cost of transportation more affordable to the public (Woodward et al., *Electric vehicles Setting a course for 2030* 2020). Lastly, Electric Vehicles allow many other primary energy sources such as nuclear, wind, solar etc. to be used in their electric form. Thus, allowing more flexibility and less interdependence on vested energy sources.

V Charging Stations by County, based on the most relevant factors

Purpose of the Report

To predict the number of EV charging stations by county based on the socio-economic, geography, census, and cost of living index. We want to analyze how these features indirectly affect demands and consumer behaviors in electric vehicle charging stations.

Methodologies

This project utilizes three different machine learning models. The first model is linear regression which includes various regression models such as OLS (Ordinary Least Square), Multiple Linear Regression and Polynomial Linear Regression. The second model is decision tree which includes the model evaluation and search function. The last model is kNN regressor and recommender which include model building and prediction system.

End to End Procedure

Data Collection

The dataset of EV charging stations is collected from afdc.energy.gov. This dataset has the lowest level of granularity wherein each of EV charging station uses city and zip as the identifier. The dataset needs a higher level of granularity where it can merge with other datasets while retaining enough rows to make a prediction. Hence, we aggregate the dataset to a county level using FIP as the primary to key to join with other datasets.

In addition, we take consideration of other possibilities in every dataset and add all geography identifiers – such as two-letters state abbreviation, state names, county names with or without “county”, city, zip code and FIP. As some states share the same county names as other states, we concatenate county name and state for another identifier.

Some datasets are aggregated to State level as the datasets for each county are not available for public. These datasets are gas price, electric price, cost of living index, grocery index, housing (renting) index, utilities index and other miscellaneous costs.

Once all the geography identifiers are added into the dataset, the next data collection is the features of the county. There are four factors that we consider when we are collecting the features of the county. These factors are census, socio-economic, cost of living and gas/electric prices.

Census factors consist of median housing price, median household income, total population, average commute time and married family. Median income and total population

Socio-economic factors consist of violent crime rate (violent crimes per 100k population), the accessibility of a community for health food and better environment, % of physical inactivity, and the number of people who drive alone. Socio-economic factors are important to describe the consumer behaviors as well as living conditions.

We add cost of living index as a measure relative cost of living and differences in the price of goods and services. In addition to cost-of-living index, we also add other measures that are connected to the necessities such as grocery, housing, and utilities.

Gas and electric prices are added into the factors as these two factors affect consumer behaviors in purchasing electric vehicles or other clean energy vehicles.

Data Ingestion and Cleaning

Our dataset's records are based on county level, originally, we have 3212 records and 30 features. All the data cleaning and wrangling has been done within python.

First step is to check the NaN value and fill in them. For instance, we fill in the na value of the median real estate with the median value of real estate per each state.

Unfortunately, we also have to drop all of the records about Puerto Rico, since that state is missing most of the information from different perspectives. We've also dropped overlapped columns, and rename the rest columns in order to make the following data analysis process easier.

Before checking the correlation heatmap, we have state names and county names, ev numbers which means the total number of ev stations in each county, it also serves as our dependent variable for the modeling. We also have a bunch of demographic features and socio- economic features serve as independent variables.

Data Visualizations

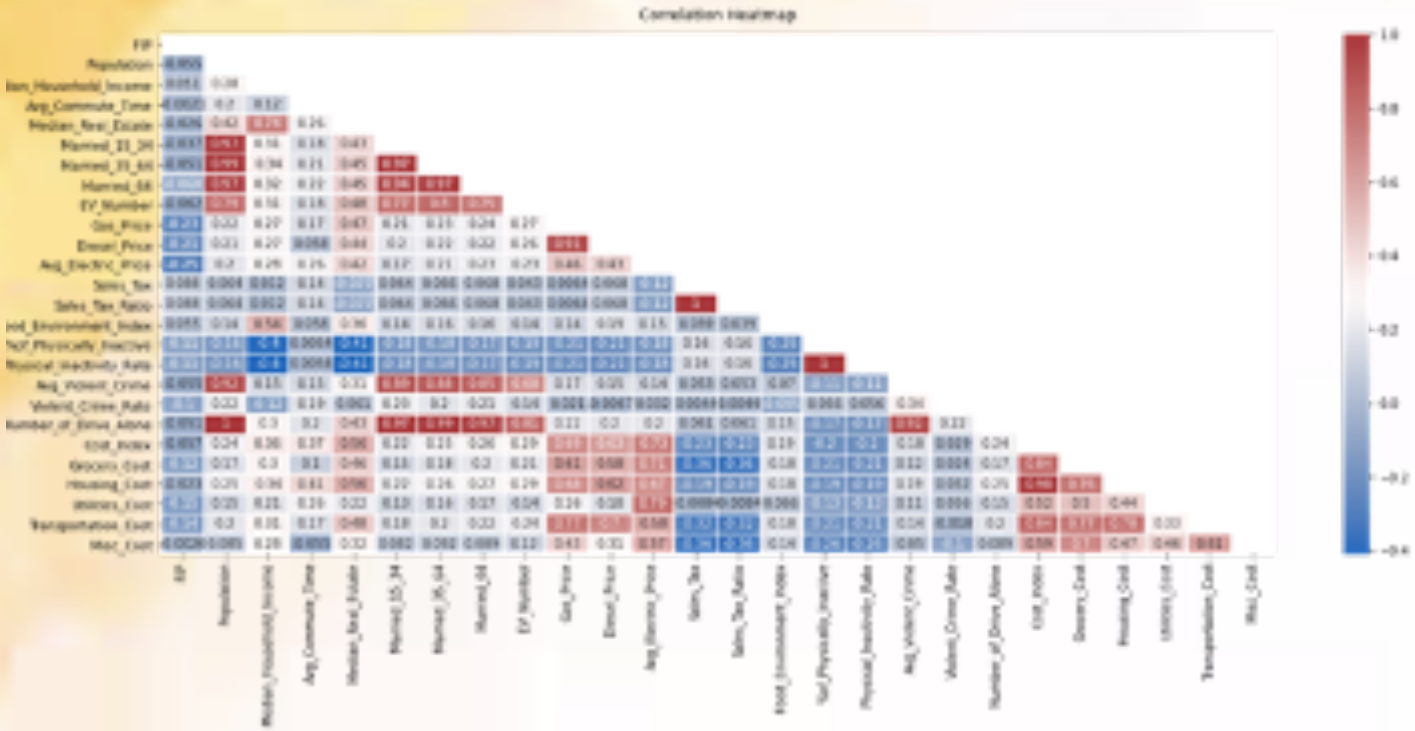
Correlation Matrix

WEused correlation matrix heatmap to observe the correlation coefficient and to see which features have the highest correlation among other features. However, we need to delete any variable with correlation above 0.7 or below -0.7 to avoid multi-collinearity.

Multi-collinearity reduces the statistical power of the regression model; hence, it may affect the final prediction model. Therefore, we eliminated any multi-collinear variables and kept the other half of the pair.

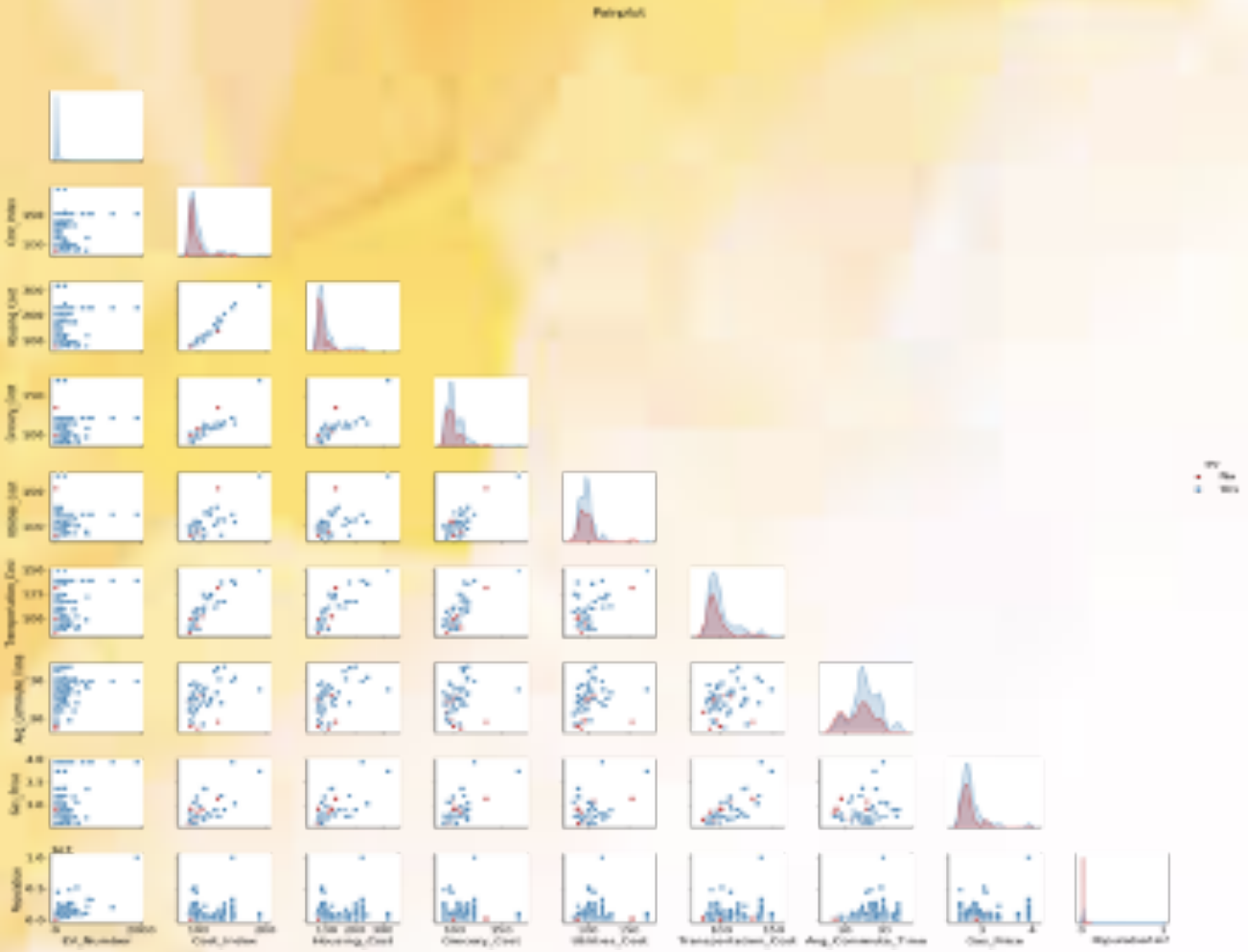
We deleted the following variables with the highest multicollinearity:

- Sales Tax
- % of physical inactivity
- Married family of 64 years old and above



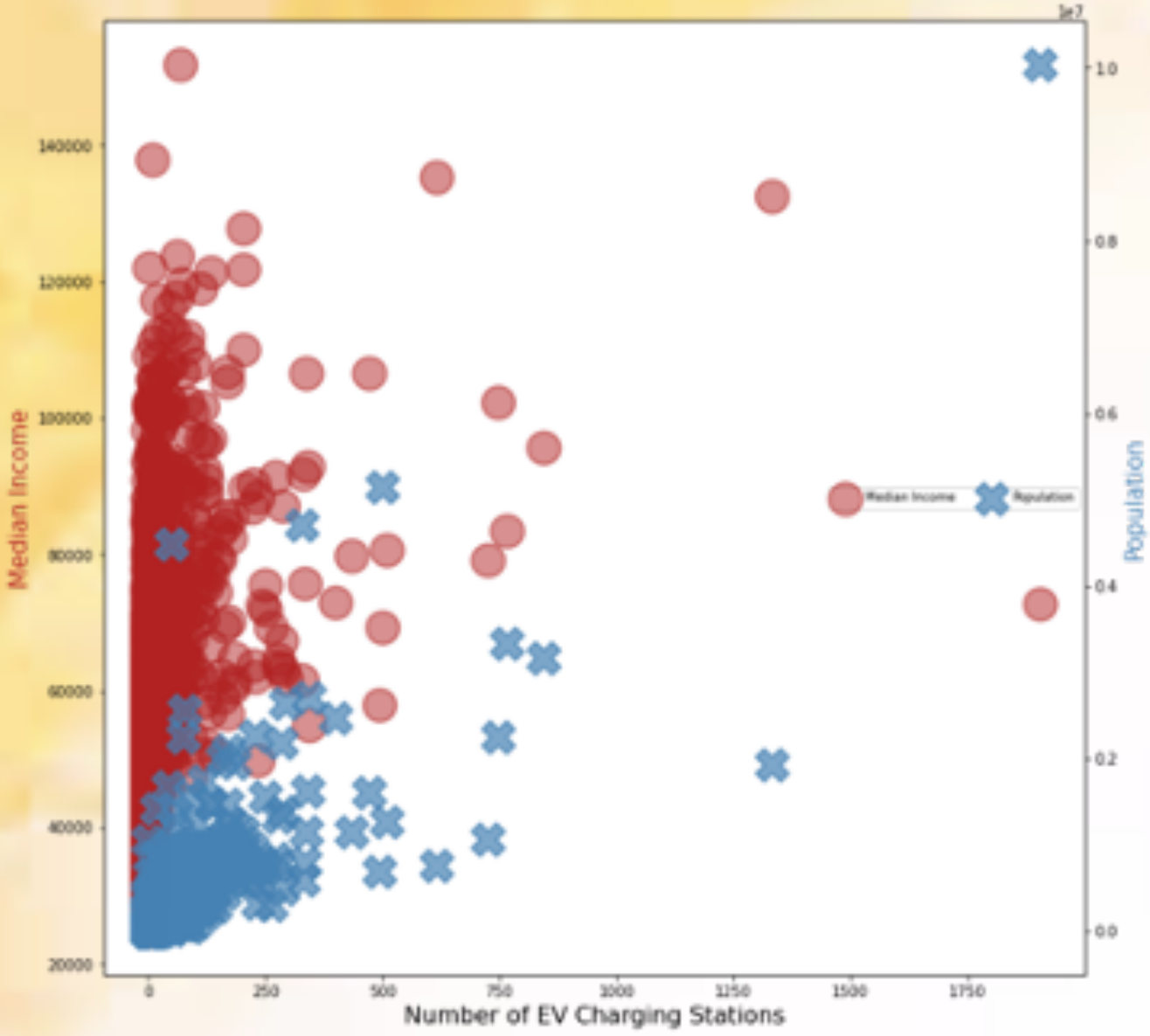
Pairplot

Pairplot is a good visualization to show the relationship between one feature with other features. It is also a good visualization to show the type of regression model to use for the model. The graph shows that the distribution of our dataset is clustering around the low number of EV stations.



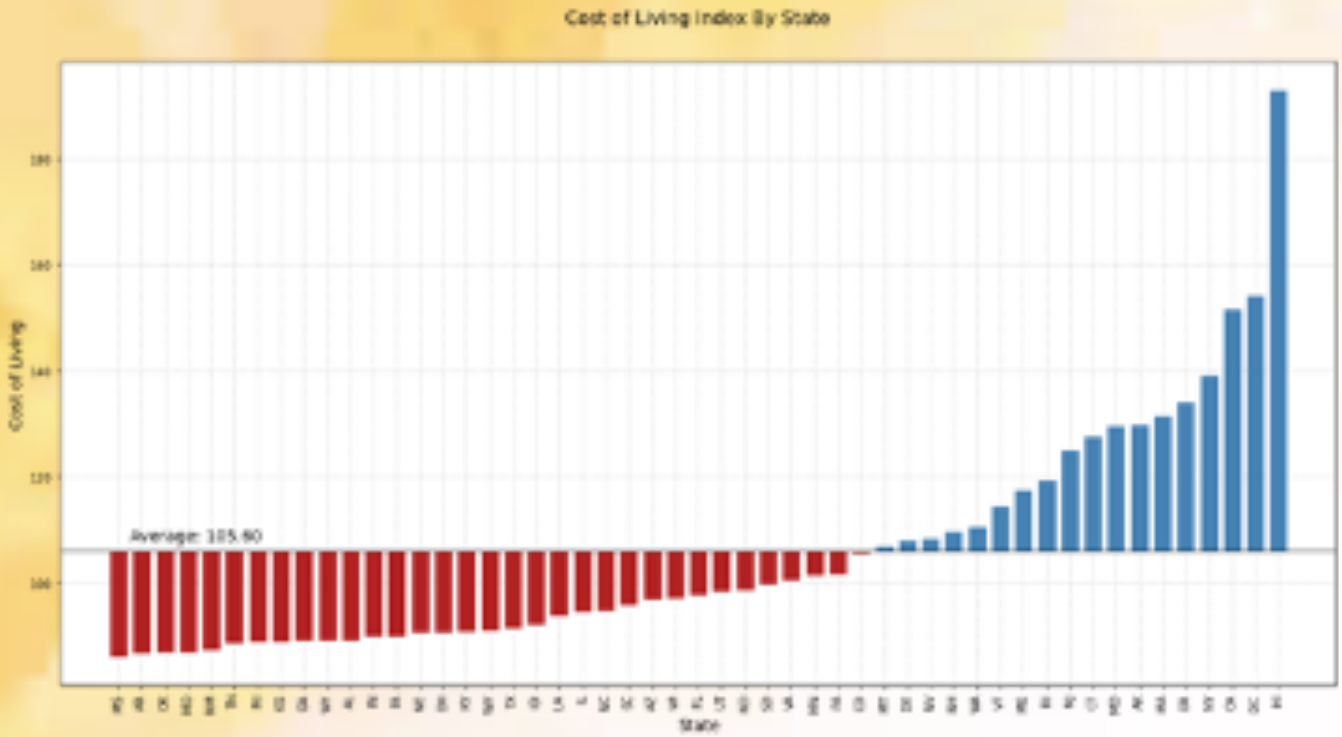
Scatter Plot

Population and median income are the most important variables that all our models will choose as independent variables. Our goal is to see the distribution of EV chargers when compared to these two variables. The graph shows that the distribution of the dataset gathers on the lower number of EV charging stations.



Diverging Chart

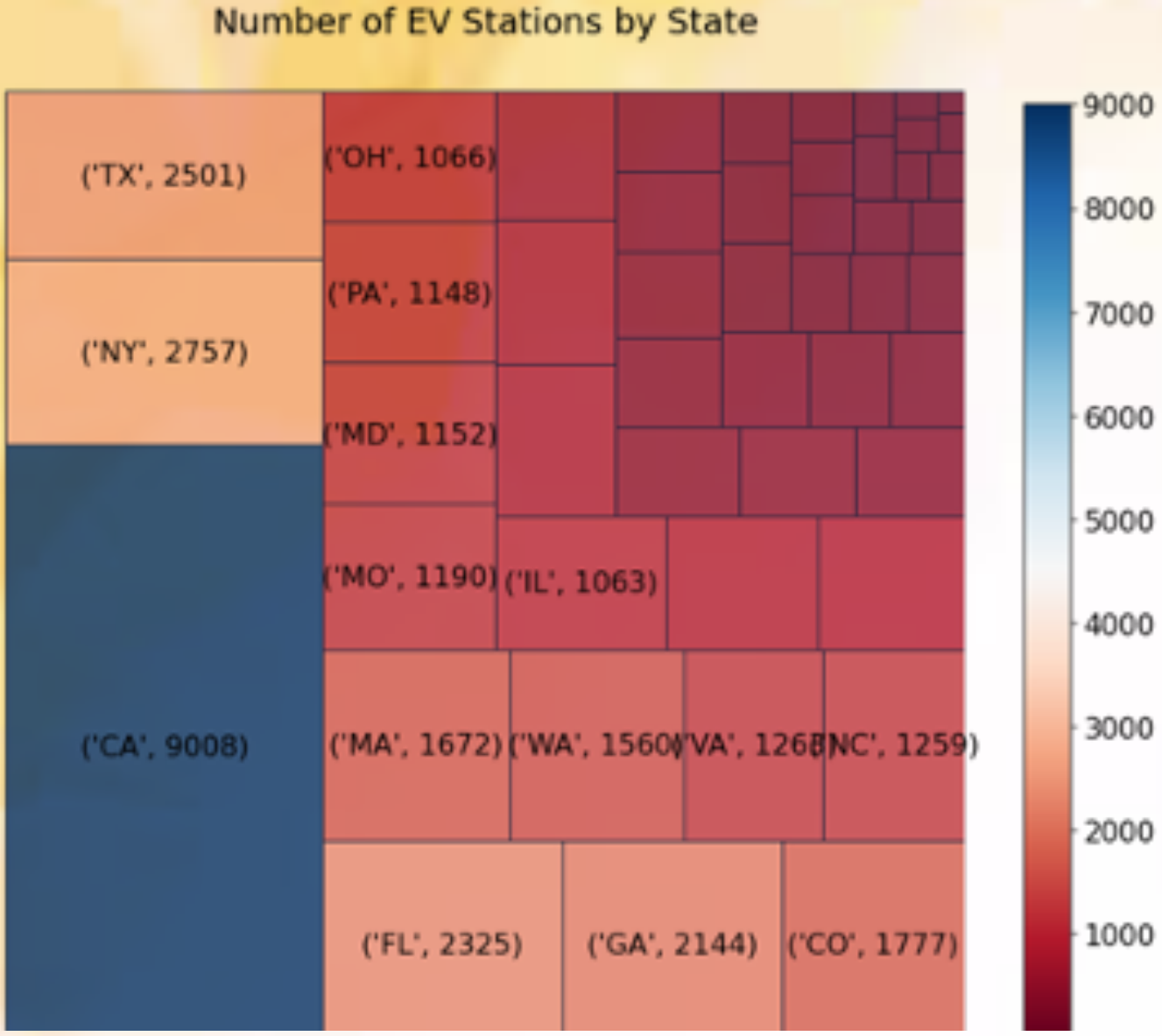
Cost of living index is a measure relative cost of living and differences in the price of goods and service. It is a measure to make a comparison between the cost of living in one state and other states. We use diverging chart to show which states have higher or lower index when compared to the national average.



Tree Map

Most of the EV charging stations are in California and followed by New York States. However, the difference between these two states is about 6000 charging stations. Hence, tree map visualization is used to standardize the disparity between the highest and lowest values.

The last visualization is to show the number of EV charging stations aggregated by states. You can tell that California is the highest and followed by NY.



Splitting Dataset into X (Variables) and Y (Target)

Ensuring that the analytics process is clear and clean, we helped the team split the dataset into X and Y. X is an independent variable and Y is the dependent variable, we are aiming to investigate within one county, how many EV stations it should have based on socio-economic features as well as other factors. So, we set the EV station number as the target variable.

Machine Learning Models

1. Linear Regression

For linear regression, we ran four distinct models. And with respect to our methodologies, we are going to go over three different regression methods. Them being, OST (which is Simple Linear via Statmodels.api) , Multiple Linear Regression (via scikit-learn) & Polynomial Linear Regression.

I. Ordinary Least Square (OLS) Regression using Statsmodels.apwe(All Independent Variables in X)

```
#OLS Regression to check for high p-values (>0.05)
import statsmodels.api as sm
model = sm.OLS(Y, X)
model = model.fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              EV_Number   R-squared (uncentered):                   0.689
Model:                            OLS   Adj. R-squared (uncentered):              0.687
Method:                 Least Squares   F-statistic:                              362.3
Date:                Mon, 03 May 2021   Prob (F-statistic):                        0.00
Time:                        18:20:18   Log-Likelihood:                         -15661.
No. Observations:                3131   AIC:                                  3.136e+04
Df Residuals:                    3112   BIC:                                  3.147e+04
Df Model:                          19
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Population     -165.6369     12.265    -13.505      0.000    -189.685    -141.589
Median_Household_Income      -6.2029      1.212     -5.117      0.000      -8.580      -3.826
Avg_Commute_Time      -1.5188      0.815     -1.864      0.062      -3.116       0.079
Median_Real_Estate      10.0318      1.265      7.932      0.000       7.552      12.512
Married_15_34     -15.4296      2.918     -5.303      0.000     -21.135      -9.724
Married_35_64      -7.7633      8.606     -0.970      0.332     -23.461       7.934
Gas_Price       6.1599      1.261      4.963      0.000       3.727       8.593
Avg_Electric_Price       0.5622      1.421      0.396      0.692      -2.225       3.349
Sales_Tax_Ratio       0.3751      0.836      0.449      0.654      -1.264       2.014
Food_Environment_Index      -1.0456      0.779     -1.343      0.179      -2.572       0.481
Physical_Inactivity_Rate       0.0403      0.748      0.054      0.957      -1.426       1.507
Avg_Violent_Crime     -13.8372      2.445     -5.660      0.000     -18.631      -9.044
Violent_Crime_Rate       0.5347      0.754      0.710      0.478      -0.943       2.012
Number_of_Drive_Alone     248.1036     11.779     21.063      0.000     225.007     271.200
Cost_Index      -7.4713      7.440     -1.004      0.315     -22.059       7.116
Grocery_Cost      -1.7867      1.653     -1.081      0.280      -5.028       1.455
Housing_Cost       8.0305      5.569      1.442      0.149      -2.890      18.951
Utilities_Cost       0.5157      1.274      0.405      0.686      -1.982       3.013
Transportation_Cost      -0.8900      1.957     -0.455      0.649      -4.728       2.948
==============================================================================
Omnibus:                     4648.458   Durbin-Watson:                   1.457
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          7332862.291
Skew:                           8.338   Prob(JB):                         0.00
Kurtosis:                     239.696   Cond. No.                         64.2
==============================================================================
```

In this method we have X as our independent Variables and Y as the Dependent Variable (X and Y description is given in the note below). In this first OLS Regression, we included all of the X variables and found a lot of them to have an extremely high p-value. As a result, the model gave an output with an R-Square of 0.689. The R-Square isn't in the high range hence proving the model to be mediocre.

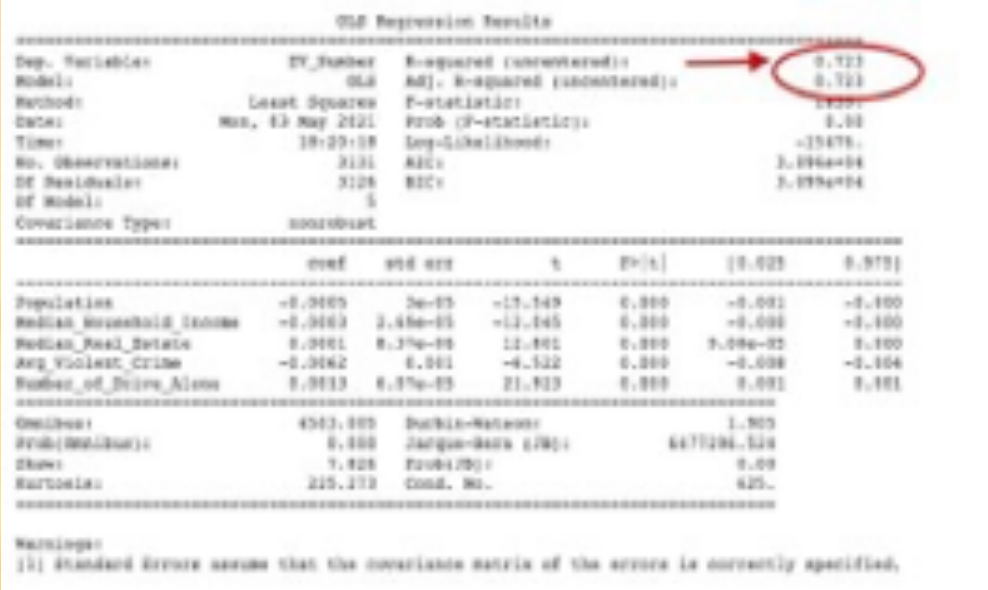
II. Ordinary Least Square (OLS) Regression using Statsmodels.apwe(Dropped Variables with P-Values > 0.05)

The output of the previous OLS Regression gave out high p-values for 'Avg_Commute_Time', 'Married_35_64', 'Avg_Electric_Price', 'Sales_Tax_Ratio', 'Food-Enviornment_Index', 'Physical_Inactivity_Rate', 'Violent_Crime_Rate', 'Number_of_Drive_Alone', 'Cost_Index', 'Grocery_Cost', 'Housing_Cost', 'Utilities_Cost', and 'Transportation_Cost'.

Therefore, it was decided to drop them and re-run the regression with the remaining independent variables. Here X1 is our Independent Variable with the dropped features. This time the Coefficient of Determinants (R-Square) gave out a value of 0.723.

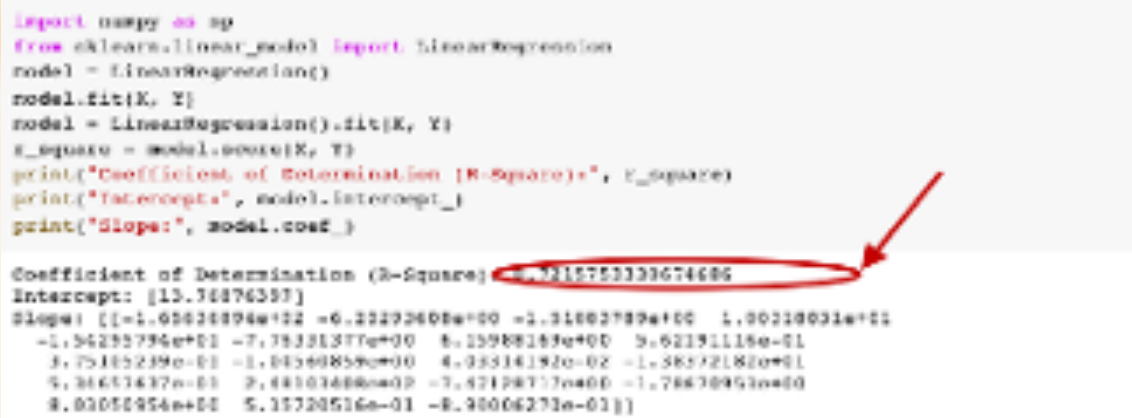
```
[41] #Creating X1 after dropping variables
     X1 = df_ev.loc[:,['Population', 'Median_Household_Income', 'Median_Real_Estate', 'Avg_Violent_Crime','Number_of_Drive_Alone',]]
```

```
#OLS Regression Result with dropped variables
import statsmodels.api as sm
model = sm.OLS(Y, X1)
model = model.fit()
print(model.summary())
```



III. Multiple Linear Regression (Using Scikit-Learn)

In this method, we imported the Linear Regression module from the Scikit-Learn package. Here, we used X as our independent Variables and Y as the Dependent Variable (X and Y description is given in the note below). After the model was run, it gave a Coefficient of Determinant output (R-Square) of 0.7216. This is even lower than the previous two OLS Regression Models



IV. Multiple Polynomial Linear Regression (Using Scikit-Learn)

Lastly, we ran a polynomial linear regression. The Coefficient of Determinants (R-Square) is the highest for this model to be north of 0.95.

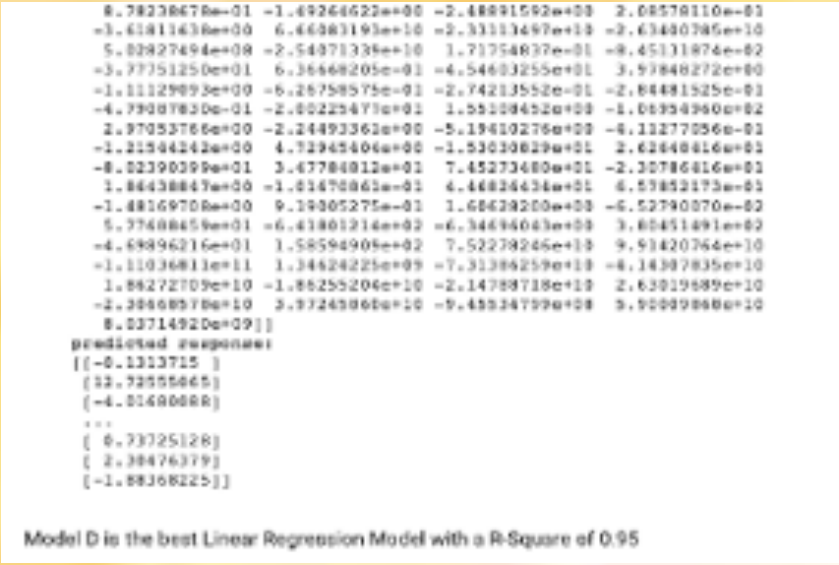
Hence, out of all the linear regression models, the polynomial linear regression model proved to be the "best with the highest R-Square" value of 0.95.

That being said, Polynomial Regression is better at fitting the data than linear regression.

Also, due to better-fitting, the RMSE of Polynomial Regression is way lower than that of

```
# Step 1: Import packages
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures

# Step 2b: Transform input data
X_ = PolynomialFeatures(degree=2, include_bias=False).fit_transform(X)

# Step 3: Create a model and fit it

model = LinearRegression().fit(X_, Y)

# Step 4: Get results
r_sq = model.score(X_, Y)
intercept, coefficients = model.intercept_, model.coef_

# Step 5: Predict
y_pred = model.predict(X_)

print('coefficient of determination (R-Square):', r_sq)
print('intercept:', intercept)
print('coefficients:', coefficients, sep='\n')
print('predicted response:', y_pred, sep='\n')
```

```
coefficient of determination (R-Square): 0.9586071288051145
intercept: [-7.65493466e+09]
coefficients:
[[-1.68617568e+02  2.60846891e+00  6.40861418e+09  5.35294981e+00
  -2.54271668e+01  4.12383012e+01  9.96231891e+00 -2.51048451e+10
   5.03802359e+09 -6.32242519e+00 -1.69583758e+00  6.82897628e+01
  -6.68210688e+02  1.39370438e+02 -1.50151160e+10  3.37586743e+10
  -1.93719953e+10  5.56381148e+09  9.06583913e+00  3.54139863e+01
   6.12491698e+02  4.87233142e+01 -4.74315803e+01 -7.09436955e+01
   2.23855862e+02  6.63367138e+01 -5.25987104e+01 -3.17191391e+01
   3.13477137e+01  6.07589823e+01  3.79179963e+01 -6.41573156e+01
  -2.13787162e+02  8.08160125e+02 -4.62984402e+01 -6.34798579e+02
  -6.38525938e+01 -2.01937503e+02  1.25985963e+00  8.91966125e-01
  -5.91893155e+00 -1.58569576e+01  1.52771455e+01  3.79233360e-02
   2.08810337e+02 -1.79139163e+00 -1.03612603e+00 -1.41312272e+00
```



Performance Analysis for the Best Regression Model: -

Therefore, from the above findings with respect to the linear regression, we can infer that the Multiple Polynomial Linear Regression is the best regression model. The performance of the model is high with a R-Square value of 0.9504. Also, the polynomial linear regression is said to give the best approximation with regards to the relationship amongst the independent (X) and dependent variables (Y).

**Note:-*

1. ***Y = number of EV Charging Stations whereas***

X = independent variables including

'Population',

'Median_Household_Income',

'Avg_Commute_Time',

'Median_Real_Estate',

'Married_15_34',

'Married_35_64',

'Gas_Price',

'Avg_Electric_Price',

'Sales_Tax_Ratio',

'Food_Environment_Index'.

'*Physical_Inactivity_Rate*',

'*Avg_Violent_Crime*',

'*Violent_Crime_Rate*',

'*Number_of_People_Driving_Alone*',

'*Cost_Index*',

'*Grocery_Cost*',

'*Housing_Cost*',

'*Utilities_Cost*',

'Transportation_Cost'

2. Decision Tree

Pre-Modeling

Our model 2 is decision tree, the reason why we chose decision tree is because we can not only predict a specific number of ev stations in a county, but also be able to find out if one county with specific socio-economic features will have above average ev stations number or below average EV stations number.

At the beginning of this model, I've also checked the overall regression result before doing the modeling. The r-squared is around 73 %, which means our dataset at the beginning is not good enough but also not bad. Dropping the variables with high p-value, and also digging into the question in order to finally drop the columns that won't contribute to the prediction. The R-squared has been slightly changed, it also indicated that WEmade the right choice to drop some features, since they don't have a huge effect on dependent variable.

Above average or below average

Setting an ev threshold for the number of ev stations is my next step, if one county has more than 15 ev stations we will say it is 0, in the opposite way it will be 1. We decided to use 15 because the average ev station number is around 13.768. We can just round it up to 13 or 14 but since as time passes by, we will expect more EV stations in a single county, so that We set the bar as 15 which is a little bit higher than the average.

Standardize

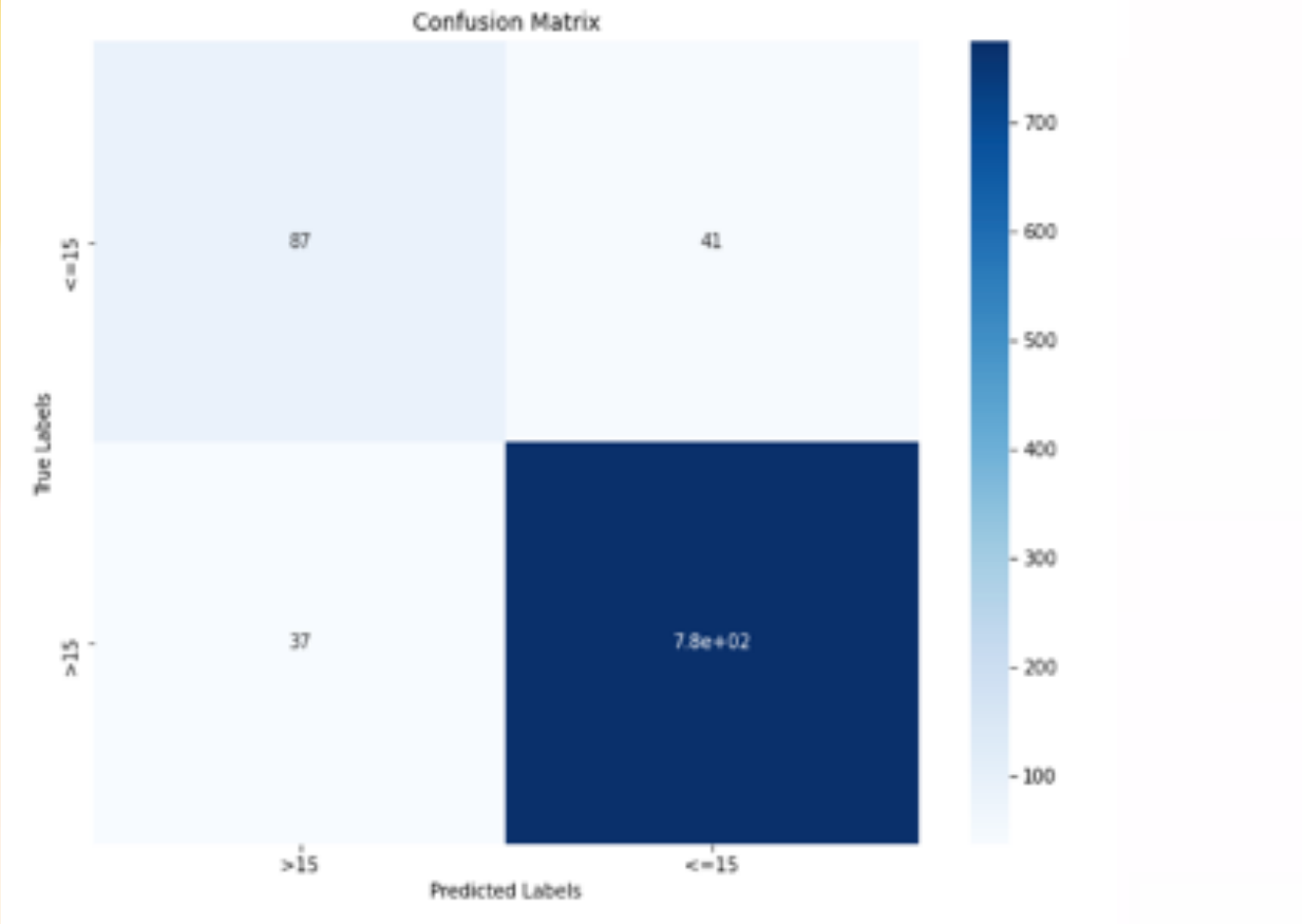
We also added one more scaling step to standardize the dataset, but the model performance is not good enough, so we changed to use bin to standardize the dataset. Binning all of the ten features one by one in order to apply for the decision tree model better. Based on the percentage of quartiles, we leveled each feature to 4 levels.

Decision tree model tuning

The first step of model tuning is split test and train dataset, we decided to separate by 70%. Also imported visualization package and modeling package before officially training the dataset. The training process is pretty straightforward. We used the regular hyperparameter as follows: max_depth = 10, random_state = 101, max_features = None, and min_samples_leaf = 15

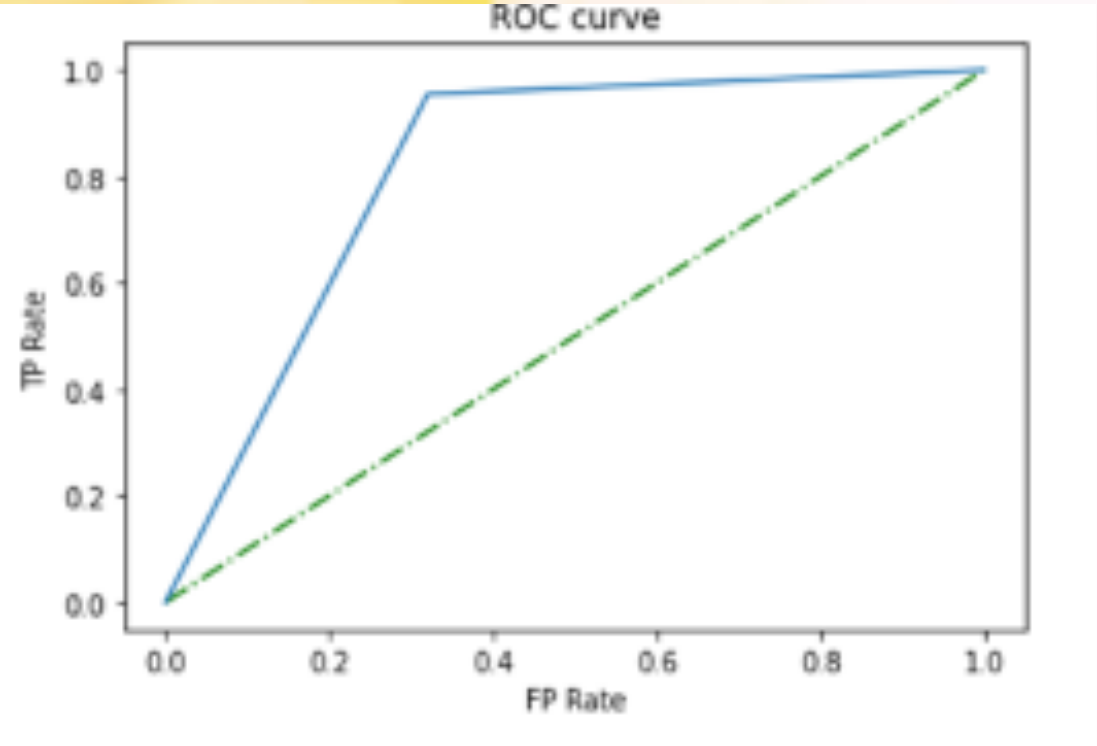
Evaluation

Accuracy is the number of good predictions over the number of predictions. In this model, we got 91% accuracy. Even though this is a high accuracy, we still cannot generate a conclusion, that if it is a good model to be chosen or not. So let's consider other measurements. The confusion matrix can help us to understand more about our model performance.



True Negative is 780, meaning 780 negative class data points were correctly classified by the model, in the same way we can say that 87 positive class data points were correctly classified by the model. False positive and false negative are small. This turned out to be a pretty decent classifier for our dataset considering the relatively larger number of true positive and true negative values.

In addition, the auc is around 81.3 which is not good nor bad. The roc curve graph also indicates that our model performance is good to predict that if a county will have ev stations above the average (15) or below average.



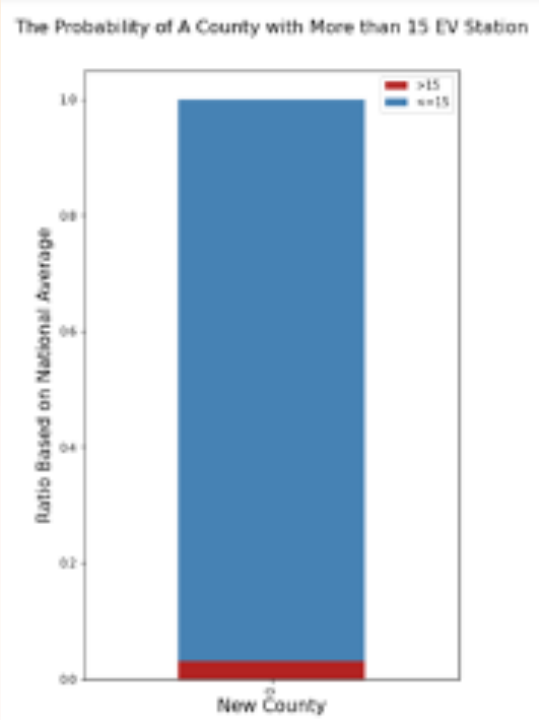
Auto tuning

For more accurate prediction, we created the auto tuning to find the best hyperparameters for the prediction. Then find out the best tree model based on accuracy and other hyperparameters. Such as Recall, since it will ensure that true positive rate will be consistent across all threshold values and can see how Recall will behave with respect to the desired threshold value.

	Split Criteria	Minimum Sample Split	Minimum Sample Leaf	Maximum Depth	Accuracy	Recall	Precision	F1Score
0	entropy	10	10	4	0.927660	0.980559	0.939464	0.959572
1	entropy	3	10	4	0.927660	0.980559	0.939464	0.959572
2	entropy	10	7	4	0.927660	0.980559	0.939464	0.959572
3	entropy	10	10	6	0.929787	0.963548	0.956574	0.960048
4	gini	10	10	4	0.929787	0.974484	0.946871	0.960479
5	gini	3	10	4	0.929787	0.974484	0.946871	0.960479
6	gini	10	7	4	0.930851	0.978129	0.944836	0.961194
7	gini	10	10	6	0.931915	0.963548	0.958888	0.961212

Search function

Furthermore, we use the best tree to create a search function. Overall, it functions as the following steps. People need to put the features list below into the function. All the features have already been leveled, only input with the decimal numbers or integer numbers. This function will automatically tell you how much is the probability that the county has these kinds of features will have ev stations number above average or below average.



Overall, the decision tree model predicts the number of EV station numbers in terms of range, as we mentioned before. It aims to help the government to predict if one county in the future, for example, urbanizationing, how the total number of EV stations should be.

Conclusion

Based on our decision tree model result, it is a good model to predict the range of number of EV stations in terms of specific socio-economic. The search function will be useful when the government wants to predict the number of ev could be in a county based on 10 socio-economic and census features. Overall speaking, the model is clear and meaningful, but we all know that there is always space for prediction models to be improved.

In summary, the decision tree is a good model to predict the range of number of ev stations for the whole project.

3. K-Nearest Neighbor and Recommender

kNN regression uses "feature similarity" to predict the values of any new data points to build a prediction model. Since our objective is to predict the number of EV charging station in a county level, hence, kNN is the perfect model to look for the resemblance within a particular county for another county.

After regression model, we build a recommender system to predict the approximate average number of EV charging stations that a county should have based on any existing county.

Regression

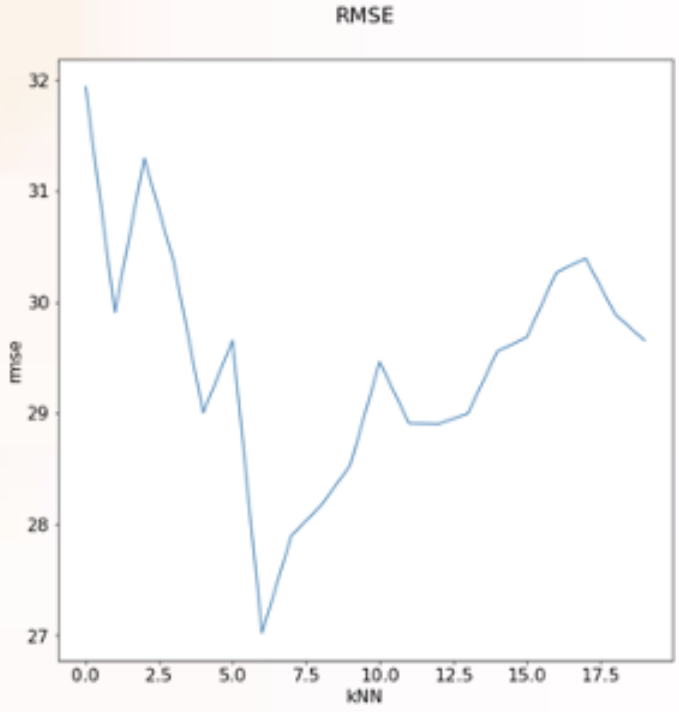
We used OLS (ordinary least square) regression results to analyze the statistical significance at the alpha of 5%. The regression results help us to remove all features that have no relationship with the number of EV charging stations.

Once 9 features are removed, the adjusted R^2 of the model dropped by 0.02. 71.9% R^2 means that 71.9% of the EV charging stations in the dataset can be explained by 10 variables that we selected. We can conclude that this model is perfect despite having lower number of features.

The scatter plot graph shows that the prediction has the least error when the number of EV charger is low. On the other hand, the model has difficulty to build a prediction when the number of EV charging stations is high.

kNN Regressor

Based on the selected 10 features, we fit our model and analyze the best model based on the RMSE value. kNN regression model is evaluated with RMSE (Root Mean Squared Error) because we want to quantify the error in our prediction model and analyze how close the predicted value when compared with the real value. It shows that the model has the lowest RMSE when K equals to 6. We also evaluate the model with model score and the score of our model is about 65% which is pretty good to build a recommender system



kNN Recommender System

The main idea behind this recommender system is to return counties with a similar feature that the user inputs. The model allows the user to input county, state, and the number of returned values. However, we need to take a consideration of human error and the lack of knowledge in geography. For example, the user might mistype the name of the county, inputs the name of the county with camel case and get confused with another county in another state. To fix the human error, we used fuzzywuzzy to match the string values by extracting the index of that value.

We used sparse matrices to compress all my features that the model fits. The recommender system is built with n_neighbors equal to 6 based on the kNN regressor model with the lowest RMSE score. The logic behind the model is to return the prediction value and to return the number of the nearest counties that we selected.

Therefore, the county that we input into the function should not be part of the nearest counties. However, the prediction value must take an account for the county that we choose.

For example, if we search for three counties that are similar to the features in San Diego, the recommender system returns only the three closest counties in term of the features. The predicted value is the average of San Diego, Orange County, Miami-Dade County and Dallas County.

The model also takes a consideration of automation system, in which the model developer does not adjust the algorithm whenever users input a different county.

```
search('San Diego', 'CA', 3)

Selected County:  San Diego County
Searching.....

Based on the features of San Diego County (CA), the predicted number of EV charging stations that a county should have is approximately 561.0

The closest counties to San Diego County are:
```

	Area_Name	State	EV_Number
215	Orange County	CA	640
352	Miami-Dade County	FL	344
2671	Dallas County	TX	294

Observations and Conclusion

Our finding shows that linear regression, decision tree and kNN recommender systems are the best models to build a prediction model for the number of EV charging stations in the future.

A Decision tree model predicts the number of EV charging stations using the national average threshold as the basis of the prediction model. The decision tree model has a particular strength to predict based on a new feature based on the existing dataset.

A kNN model, on the other hand, predicts the number of EV stations using the features of the existing counties as the basis of the prediction model. Therefore, the kNN model is very effective to make a prediction if there are similar features or factors to compare with.

Our findings conclude that the decision tree model is a better predictor to predict the number of EV charging stations in the future.

Bibliography

Shahriar, S., Osman, A. H., & Nijim, M. (2020). (rep.). *Machine Learning Approaches for EV Charging Behavior: A Review* (pp. 1–14). The Institute of Electrical and Electronics Engineers.

Woodward, M., Waltn, B., Hamilton, J., Alberts, G., Fullerton-Smith, S., Day, E., & Ringrow, J. (2020, July 28). *Electric vehicles Setting a course for 2030*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/focus/future-of-mobility/electric-vehicle-trends-2030.html.

Dataset

Geography

FIP & ZIP: https://data.world/niccolley/us-zipcode-to-county-state

Census

Median Real estate Price: https://cdn.nar.realtor/sites/default/files/documents/2020-q3-county-median-home-prices-by-price-01-06-2020.pdf

Average commute time https://www.census.gov/search-results.html?q=Average+Commute+Time+Census&page=1&stateGeo=none&searchtype=web&cssp=SERP

Population and Income https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Married Household Family: https://data.census.gov/cedsci/table?q=married%20household%20family&tid=DECENNIALAS2010.PCT13

Gas/Electric Price

Gas Price https://gasprices.aaa.com/state-gas-price-averages/

Electricity Retail Price By state https://www.eia.gov/electricity/state/

Socio-economic factors

50 Datasets of food index, violence rate, % of physical inactivity, # of drivers who drive alone https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-factors/social-and-economic-factors/community-safety/violent-crime-rate

Cost of Living Index

Cost of living index https://worldpopulationreview.com/state-rankings/cost-of-living-index-by-state

Cost of living index DC https://www.bestplaces.net/cost_of_living/city/district_of_columbia/washington

Number of EV chargers