Bryan Lihardinata
BSAN6070 – Machine Learning

Link:
https://colab.research.google.com/drive/1j5NR22HoT5fSOQUNeqLkSVSGcnsLOG9t#scrollTo=6aF7_WQQHs3w
https://github.com/blihardinata/EV_charging_station/blob/main/Bryan_Lihardinata_EV_Charging_Stations.ipynb

Objective: To predict the number of EV charging stations that a county should have based on socio-economic features as well as other factors.

Data Cleaning:

Most of the data collections and merging are executed using Excel and Tableau Prep.

Dataset:

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 are the most two important features in the urban level.

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 Visualization

Correlation Matrix

I used 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, I eliminated any multi-collinear variables and kept the other half of the pair.

I 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.

<u>kNN Modelling</u>

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.

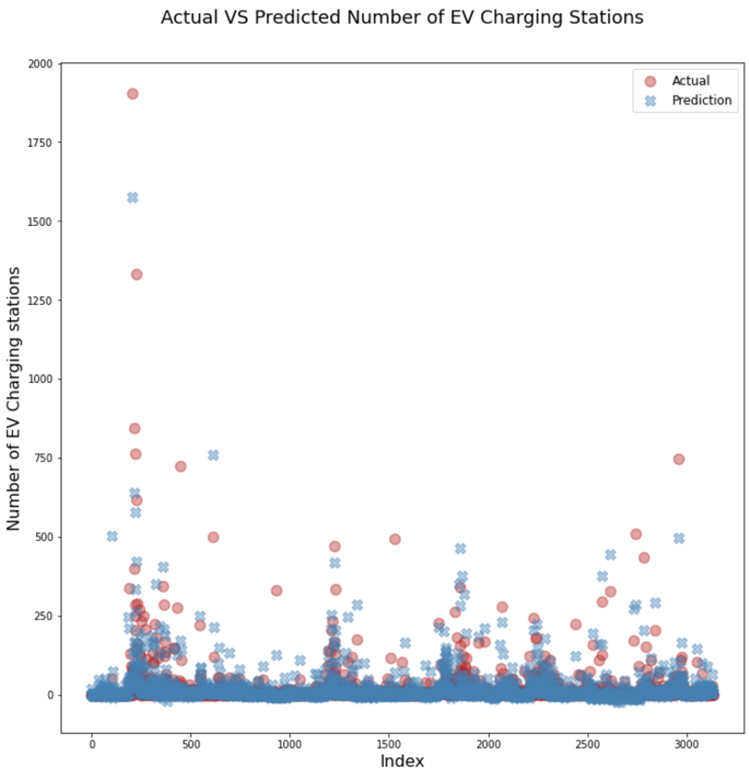
Regression

I 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.

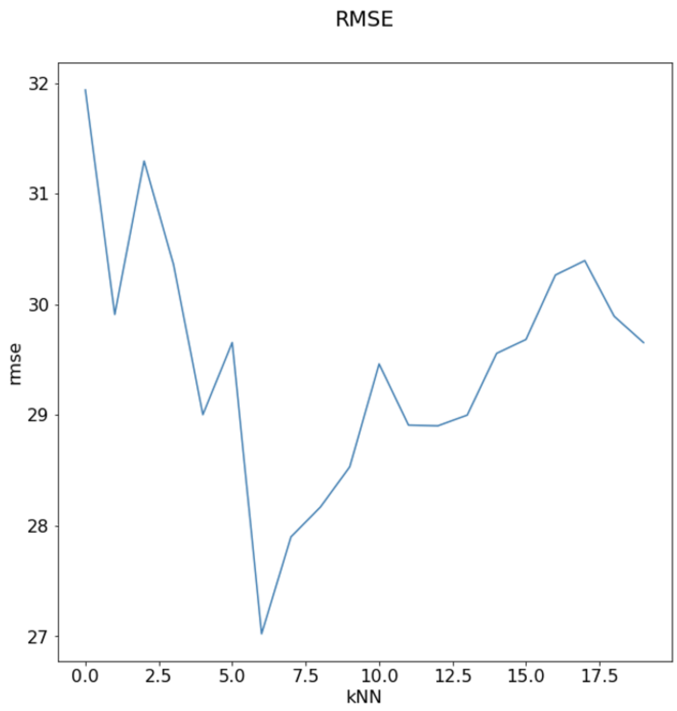
(I will revise my model with robust linear regression as advised in the class)



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.

(Side note: the RMSE scores and kNN nearest neighbors fluctuate every time the codes are executed.)



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, I used fuzzywuzzy to match the string values by extracting the index of that value.

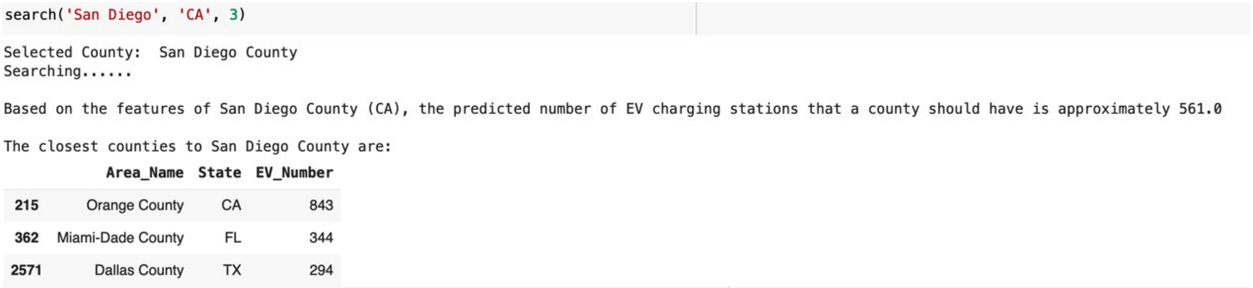
I used sparse matrices to compress all my features that the model fits.

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 chose.

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.



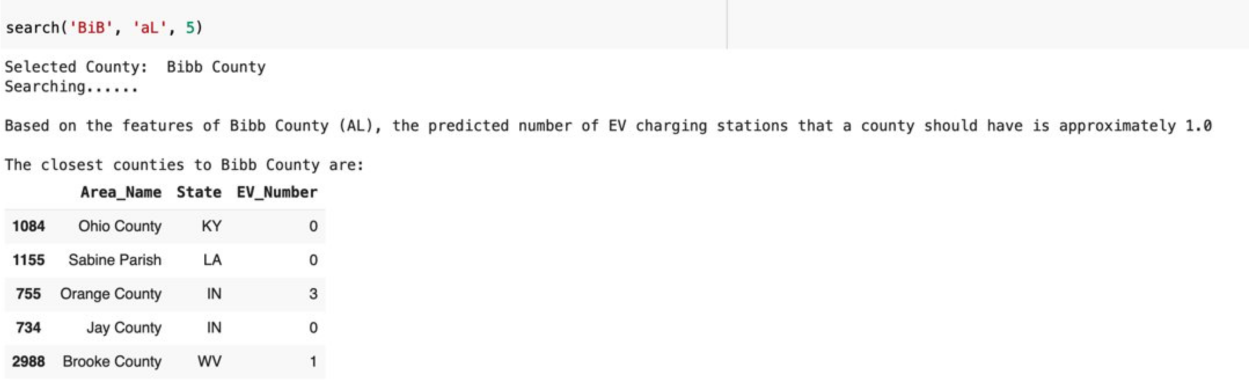
```
search('San Diego', 'CA', 3)
```

```
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	843
362	Miami-Dade County	FL	344
2571	Dallas County	TX	294



```
search('BiB', 'aL', 5)
```

```
Selected County:  Bibb County
Searching......

Based on the features of Bibb County (AL), the predicted number of EV charging stations that a county should have is approximately 1.0

The closest counties to Bibb County are:
```

	Area_Name	State	EV_Number
1084	Ohio County	KY	0
1155	Sabine Parish	LA	0
755	Orange County	IN	3
734	Jay County	IN	0
2988	Brooke County	WV	1

Conclusion

Based on our finding, kNN recommender system is a good model to predict the number of EV charging stations based on socio-economic factors, census, and all other factors. It returns the approximate values of prediction based on the existing counties with similar features. Therefore, the model is very effective to predict a value that the model can make a comparison. However, the model needs more adjustment if we need to add more features. Additionally, the model does not have something to compare with should we want to make a prediction based on new inputs.

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.

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

Number of EV charger https://afdc.energy.gov/fuels/electricity_locations.html#/analyze