Abstract

Can we predict the median salary of an area based on the following: COVID-19 mandate violation citation rate, COVID-19 case rate, COVID-19 unvaccinated rate, and COVID-19 death rate?

Based on the results of this project the answer is yes though not all of the features are required to answer the question. I found that the best prediction model is the KNN model and that only 2 features, COVID-19 mandate violation citation rate and COVID-19 death rate had an impact on the result.

Introduction

I wanted to see if there was a relationship and there a way to predict the median salary of an area based on different COVID-19 features. I read many articles with regards to COVID-19. Some were about religious arguments against vaccines, COVID-19 mandate violations, COVID-19 deaths amongst low income earners, etc.

All of this information got me thinking about what other relationships could be made between COVID-19 and median income and the affect of religion on vaccine adoption.

Unfortunately I was unable to ascertain a relationship between religion and vaccine adoption so decided to focus on COVID-19 and median income.

Based on the articles I read I was expecting to find a negative relationship between all of the COVID-19 features and the median income and I did for all but one feature. I was surprised to see a positive relationship between COVID-19 mandate violation citation rate and median income.

Data

Citations for COVID-19 Related Violations

(COVID-19 Locations & Demographics - LA County Department of Public Health)

This dataset details 1,639 COVID-19 related violations as of 01/28/2022.

Information includes activity date, company name, address, city, and company description. The complication with this data is that it's missing the zip code.

In order to process this dataset I summarized the number of citations per city.

This data was provided as a table on the website so I needed copy and paste it into Excel. I then saved the data as a CSV file.

COVID-19 Vaccination Rate by City

(COVID-19 Vaccine Progress Dashboard Data by ZIP Code - Datasets - California Health and Human Services Open Data Portal)

The dataset summarizes the COVID-19 Vaccinations administered by zip code as of 02/01/2022.

Information includes zip code, county, population, vaccination number. There is more data but for this project I only required the data listed. The complication with this dataset is that it's by zip code where other datasets are by city.

This data was provided as a table on the website so I needed copy and paste it into Excel. I then saved the data as a CSV file.

COVID-19 Case Rate & Death Rate by City

(COVID-19 Locations & Demographics - LA County Department of Public Health)

This dataset summarizes the total number of COVID-19 cases and deaths by city as of 01/28/2022.

Information includes City, Case Number, Case Rate, Death Number, Death Rate. The complication with this dataset is the naming of the different cities. I needed to convert these names to the official names in order to match it to other datasets.

Median Income by Los Angeles County Neighborhoods

(Median Income Ranking - Mapping L.A. - Los Angeles Times (latimes.com))

This dataset summarizes the median income by neighborhood in Los Angeles county.

This data includes Neighborhood, Ranking, and Median Salary. The complication with this dataset is some neighborhoods are not cities so I needed a way of mapping them.

This data was provided as a table on the website so I needed copy and paste it into Excel. I then saved the data as a CSV file.

Zip Code Database

(ZIP Code Database - ZIP Code List (unitedstateszipcodes.org))

This dataset maps zips codes to the primary city they belong to.

The main use of this dataset was to convert datasets at the zip code level to the appropriate city level.

Los Angeles County Neighborhoods

(Los Angeles County Neighborhoods - Datasets - UCLA Geoportal)

This dataset maps the different LA County neighborhoods to the appropriate city.

The main use of this dataset was to map the Median Income dataset to the correct city.

Cleaning Methods and Tools

In order to merge these datasets into a usable format I needed to use Excel's PowerQuery. This allowed me to bring the data into a single excel document where it could be summarized and joined to the other datasets.

Data Standardization

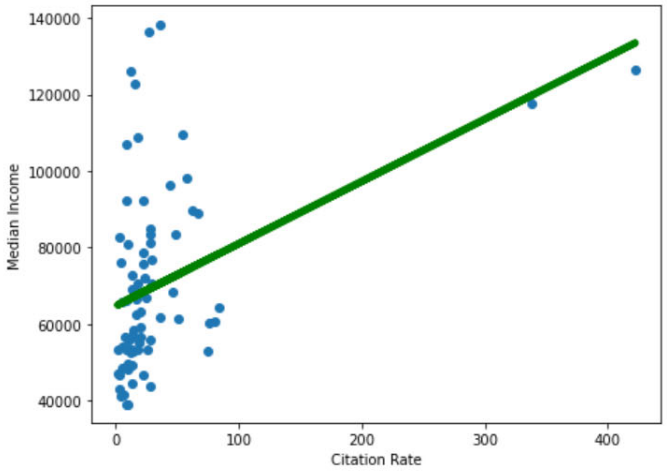
I then needed to standardize the data as each city contains different population data. I decided to standardize the counts into a rate per 100,000.

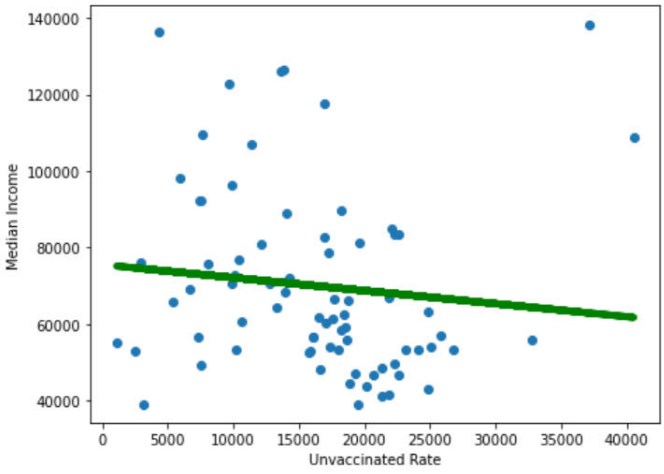
Data Impute

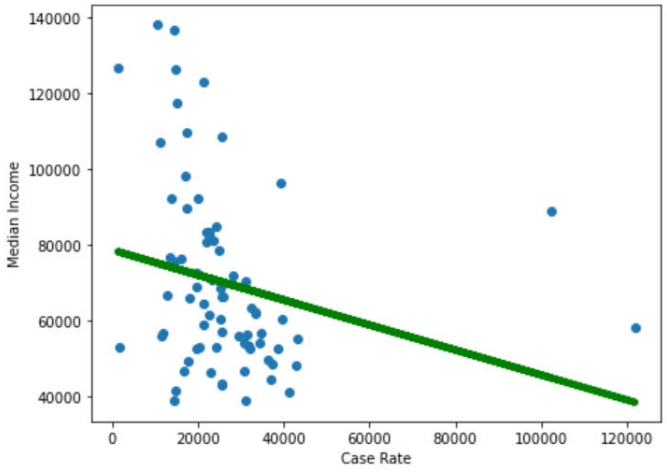
When the final dataset was imported into Python a simple imputer, using the most frequent strategy, was implemented to ensure there was no missing data.

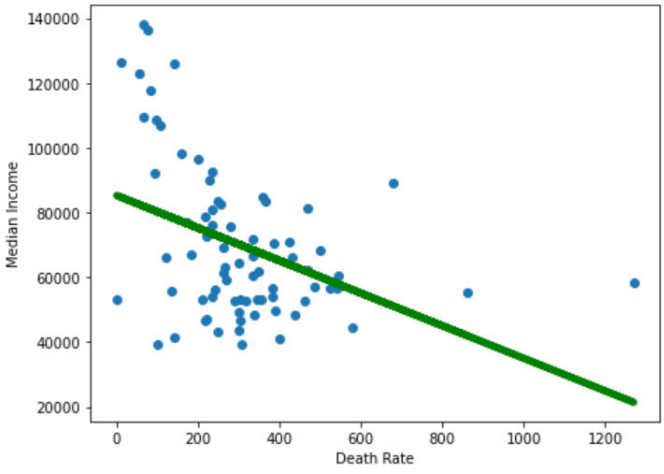
Linear Regression of Individual Features

Linear regression models for each feature was determined and mapped along with the resultant estimation line.









OLS Regression Model for All Features

An OLS Regression model was creates including all of the features.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           MedianIncome   R-squared:                       0.308
Model:                            OLS   Adj. R-squared:                  0.268
Method:                 Least Squares   F-statistic:                     7.772
Date:                Thu, 10 Mar 2022   Prob (F-statistic):           3.05e-05
Time:                        18:15:00   Log-Likelihood:                -849.08
No. Observations:                  75   AIC:                             1708.
Df Residuals:                      70   BIC:                             1720.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.364e+04   7383.113     11.328      0.000    6.89e+04    9.84e+04
CitationRate    130.5309     40.069      3.258      0.002      50.617     210.445
UnvaccinatedRate  -0.4142      0.323     -1.282      0.204      -1.059       0.230
CaseRate          0.3226      0.230      1.403      0.165      -0.136       0.781
DeathRate       -64.9141     20.511     -3.165      0.002    -105.822     -24.006
==============================================================================
Omnibus:                       11.230   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               11.658
Skew:                           0.810   Prob(JB):                      0.00294
Kurtosis:                       4.050   Cond. No.                     1.07e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.07e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

OLS Regression Model for Remaining Features

Based on the information from the above model I determined that the CaseRate and UnvaccinatedRate features did not add to the relationship as the coefficients were very close to 0 and the standard deviations included both negative and positive values.

A new OLS Regression model was created after removing those features.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           MedianIncome   R-squared:                       0.277
Model:                            OLS   Adj. R-squared:                  0.257
Method:                 Least Squares   F-statistic:                     13.82
Date:                Thu, 10 Mar 2022   Prob (F-statistic):           8.34e-06
Time:                        18:15:00   Log-Likelihood:                -850.68
No. Observations:                  75   AIC:                             1707.
Df Residuals:                      72   BIC:                             1714.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       7.813e+04   4968.219     15.725      0.000    6.82e+04     8.8e+04
CitationRate   136.4668     40.198      3.395      0.001      56.334     216.600
DeathRate      -41.4884     12.804     -3.240      0.002     -67.012     -15.964
==============================================================================
Omnibus:                       10.325   Durbin-Watson:                   1.882
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               10.291
Skew:                           0.863   Prob(JB):                      0.00583
Kurtosis:                       3.560   Cond. No.                         743.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Assumptions

I needed to make some assumptions with this analysis including:

- The influence of the companies involved in the COVID-19 mandate violation citations was predominately in the city they are located.
- I couldn't find a dataset for the number of companies per zip code so I used the population for the citation standardization even though number of companies would've been more accurate. My assumption is that the number of companies is relative to the population.
- Even though the COVID-19 mandate violation citations where only given during part of the COVID-19 time frame the affect could still be seen when looking at total deaths.
- Even though the COVID-19 vaccinations where only available during part of the COVID-19 time frame the affect could still be seen when looking at total deaths.
- The total population was determined based on the people where vaccinations are currently available (anyone who is 5 years or older.)

Methods

Process used for project

In order to answer this question I needed to perform the following steps:

- Obtain datasets
- Clean and combine datasets
- Standardize and impute the features
- Determine feature set using linear and OLS regression models
- Split dataset into training and testing datasets (80/20)
- Determine best prediction model (Linear regression, KNN, and Ridge Regularization)

Linear Regression Model - Feature selection

I used this model because I wanted to make sure there was some relationship between each of the features and the median income. In the data section I show that a relationship does exist for each feature.

OLS Regression Model

I used this model because the model can contain all of the features and provides an output that shows how each feature works together rather than individually. In the data section you can see how the UnvaccinatedRate and CaseRate features are not affecting the overall relationship. Both of them have a coefficient that is close to 0 and have a standard deviation that contains both negative and positive values.

I removed the UnvaccinatedRate and CaseRate features and reran the model. As seen in the data section the CitationRate coefficient and standard deviation didn't change much but the DeathRate coefficient and standard deviation improved.

K-nearest neighbor (KNN) and Ridge regularization Models

Since the project is about prediction the KNN and Ridge models were chosen. I needed to determine which would provide the best results. The best way to do this is to test both with different parameters (k and lambda.)

The KNN model was run for several value of k and the Ridge Regularization model was run with many values of lambda.

I the end the KNN model with a k value of 50 was the best overall model based on the training and test RMSE values.

Results

Data Split

The data was split into a training dataset (80%) and testing dataset (20%)

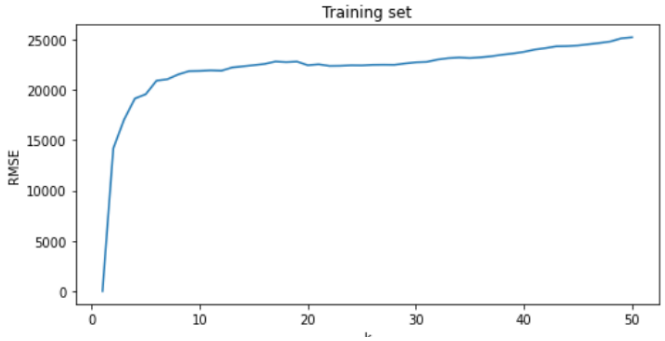
Model Determination

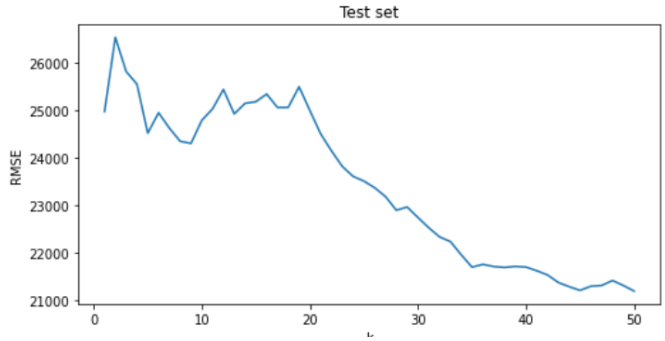
Linear regression model was created and tested with the following RMSEs:

- Train RMSE of the linear regression model is: 17907.911533629976

- Test RMSE of the linear regression model is: 30349.839365551703

K-nearest neighbor (KNN) models were created and tested using different values of k and the resulting RMSEs were plotted.

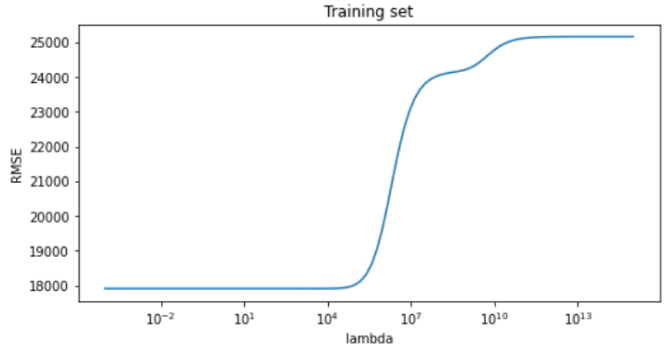


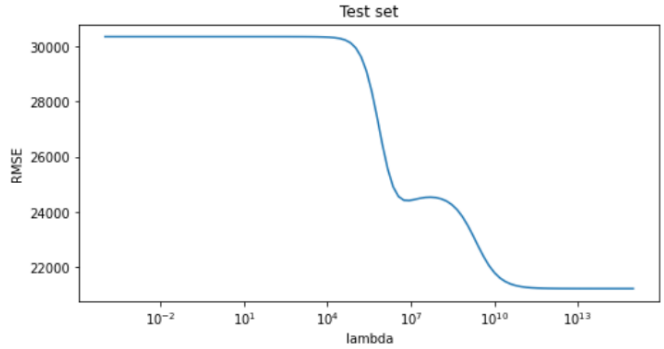


The best KNN model produced the following RMSEs:

- The minimum Validation RMSE is 25243.487673662716
- The Training RMSE for minimum Validation RMSE is 21183.635852391588
- The k value for minimum Validation RMSE is 50

Ridge regularization models were created and tested using different values of lambda and the resulting RMSEs were plotted.





The best ridge regularization model produced the following RMSEs:

- The minimum Validation RMSE is 21223.030543538753
- The Training RMSE for minimum Validation RMSE is 25156.953398912327
- The lambda for minimum Validation RMSE is 1000000000000000.0

Best Model Selection

Over all the best model was the KNN model with a k value of 50.

Conclusion

The median income of an area can be predicted based on the COVID-19 mandate violation citation rate and COVID-19 death rate. These results did partially match my intuition but no fully. I was surprised that one 2 of the 4 selected features impacted the prediction.

One question that came up was; Can you predict COVID-19 death rates based on

the following features: Religious affiliation, political affiliation, and racial diversity.

Another question was; Do the actions and messages of the church increase COVID-19 unvaccination rates along with COVID-19 death rates.

If I had more time for this project I would have increased the scope to look at the entire US rather than limiting it to just Los Angeles County. I also, would've added more features like religious affiliation, and political affiliation.