Predicting for U.S. Counties with Missing COVID-19 Reports

Wendy Hou, Bichen Kou, Emmy Phung, Lily Zhou
wh916@nyu.edu, bk2374@nyu.edu, mtp363@nyu.edu, yz6121@nyu.edu

1. Introduction

When going over Johns Hopkins' official United States COVID-19 daily reports, the team noticed that there were a lot of U.S. counties missing from the reports. For example, on 4/12/2020, 529 out of 3147 counties are missing from Johns Hopkins' daily COVID-19 report. The goal of this project is to predict the infection rate of these missing counties based on their county-specific feature data, including but not limited to population density, education, public and private healthcare information, and popular transportation methods. Furthermore, we analyzed the impacts of these factors on the spread of COVID-19. [1]

2. Data Cleaning and Preparation

The two main datasets utilized in this paper are from Johns Hopkins and the U.S. Census. The Johns Hopkins' datasets have detailed COVID-19 information on the numbers of confirmed cases, death, and recovered cases for each county. For the scope of this project, the team used the data reported on 4/12/2020, which was the most up-to-date one when we started the project. All the county features come from the U.S. Census and these features cover a broad variety of categories, forming 97 features in total (Figure 1). These features indicate each county's population, area of land, gender distribution, age distribution, employment status, means and average time of commuting, education level distribution, income distribution, and healthcare coverage. The team also purposefully steered clear of the features that could cause problematic discussions like race. The U.S. Census data was very comprehensive, except for one especially important category for this epidemic, healthcare coverage. For features related to healthcare coverage, less than 2,000 counties are reported, meaning that this dataset is missing more than 1/3 of the counties. Considering the potentially significant influences these features will have on our final prediction of infection rate for the counties, the team decided to impute the dataset.

The team first attempted to group counties based on all non-healthcare related features. We hypothesized that each cluster would have some in-group similarities about their healthcare coverage, which allowed us to impute counties' missing healthcare data based on the average of their groups. To achieve this, the team used the KMeans function from sklearn.cluster and broke the counties into 50 groups. Unfortunately, the clustering of these U.S. counties was extremely unbalanced, with cluster sizes from one county to over 1,000. Therefore, the team decided to abandon the original idea and, instead, use the nearest neighbors of these missing counties to impute their healthcare-related data. The team used the KNNImputer function from sklearn.impute and set 'n_neighbors' to one so that we only referred to the nearest neighbors of these counties. After obtaining and imputing all the feature data needed, the team generated a few more features from existing features. For example, we calculated the population density of each county by dividing 'population' over 'land area'. To limit collinearity among features, columns that were used to create new features were dropped from the feature matrix. To obtain our official target column, we gathered the number of confirmed cases from Johns Hopkins' dataset of each county on 4/12/2020 and divided it by the county's population to obtain our target column 'infection rate' for each county. Any county that was not reported on 4/12/2020 in Johns Hopkins' dataset was assigned with an infection rate of 0.0% and taken out of the dataset to form our explore set. The rest of the dataset was split into 60% training and 40% testing.

3. Prediction Results

To predict the COVID infection rates on the U.S. county-level, the team trained five models, all with hyperparameter tuning and regularization when appropriate, and chose MSE as our evaluation metric. The results of all five models are reported in table 1. The MSEs shown are relatively high considering the span of infection rate only ranges from 0.002% to 6.022% and the U.S. national infection rate is around 0.464% (calculated based on data from 1point3acres). This is possibly due to the lack of accuracy in our ground truth and the existence of outliers in our datasets. Therefore, the team switched the focus of our project from predicting a continuous number for the infection rate of each county to predicting whether the county is at high- or low-risk of COVID-19, which was a binary classification task.

Model	MSE(%)
Ridge Regression	0.031
Lasso Regression	No Convergence
Random Forest	0.021
Gradient Boosting	0.021
Neural Network	0.023

Table 1: MSE for five prediction models

[1] Github Repo: https://github.com/Emmyphung/COVID19-predict-infection-rate

To get a better sense of each feature's contribution to the prediction model, we plotted the feature coefficients, calculated by sklearn.linear_model. Interestingly, the majority of the contributing features are related to either commute time or commute type. As shown in Figure 2, 'Workers per Car, Truck or Van', a column that represents the average people in a car for each town, is the most impactive feature and it has a positive correlation with the infection rate, while features like 'Worked at Home', 'Bicycle', 'Walked', which indicate the percentage of population in a county using this means of commuting to work, have a negative correlation with the infection rate. Across all the attempted models, 'Public Transportation', which indicates the percentage of population in a county using public transportation to commute to work, is one of the most important features. Surprisingly, it has a negative correlation with our target variable according to the linear model. This might be due to feature correlation since there are about 100 features feeded to the model. To handle this problem, we ran a stepwise feature selection and picked only the statistically significant features. As shown in Figure 3 and 4, there were in total 22 features selected, among which, 'Public Transportation' is the most influential feature and has a positive correlation with the infection rate.

4. Classification Results

The results obtained from our prediction task showed that we could, to some extent, infer the infection rate of COVID-19 but obtaining the exact number was not promising. Thus, the team found an alternative approach to leverage our dataset to capture the prevalence of COVID-19 while not focusing on the exact infection rate. We shifted our focus from solving a regression problem to a binary classification one by specifying high- and low-risk counties based on their infection rate. After excluding the counties that did not report, we labeled the remaining counties as low-risk counties if their infection rate was less than the median infection rate, 0.034%, and as high-risk otherwise.

The team again trained five models with hyperparameter tuning and regularization when appropriate. We chose AUC and Recall as our evaluation metrics. We specifically evaluated our models' performance based on Recall because we believed, in this scenario, it was worse to predict a county to be low-risk when it is in fact high-risk, thus false negatives should be minimized. Here are the results of all five models:

Model	AUC	Recall
Logistic Regression	0.645	0.669
SVM	0.743	0.656
Perceptron	0.710	0.602
Gradient Boosting	0.743	0.637
Neural Network	0.680	0.528

Table 2: Classification results on test set

Based on the results, the team chose Logistic Regression and SVM to be our best models for predicting whether a county is at high or low risk. They performed the best in AUC and Recall respectively and, although SVM and Gradient Boosting performed equally well in AUC, SVM gives a higher Recall than Gradient Boosting.

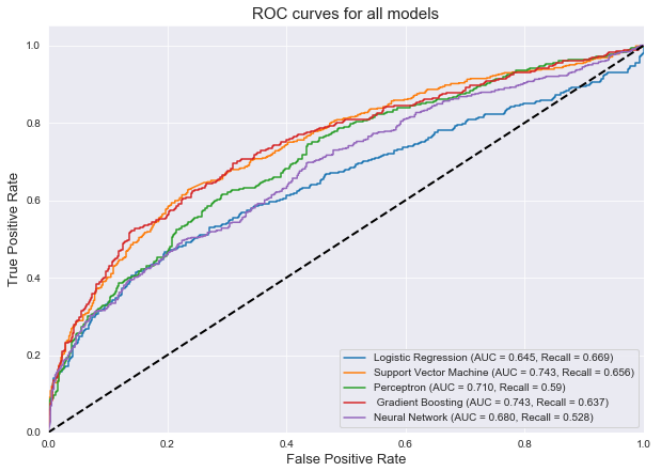

Figure 1: ROC curves of all classification models

5. Conclusion and Discussion

As mentioned previously, there are 529 counties with missing infection rate on 04/12/2020. We predict whether these counties are high-risk or low-risk counties by our best models, Logistic Regression and SVM. There are 15 counties predicted as high risk counties by Logistic Regression and 219 by SVM. There are 4 counties predicted as high-risk counties by both models and all of them are in New York. We are surprised by the drastic difference in number between our two best models. To evaluate the model performance furthermore, we calculate the infection rate based on the up-to-date data (on 05/17/2020), and classify counties as high- or low-risk counties according to the up-to date infection rate median. As shown in the table below, Logistic Regression gives a higher precision but lower recall when compared with the SVM model:

Model	Number of high-risk	Precision	Recall
Logistic Reg	14	6/14	6/67
SVM	219	32/219	32/67

Table 3: Classification results on explore set

Although we collected as much information as possible and tried different models, we still found that our models did not perform well on the tasks, prediction or classification. One of the most important reasons is that our target is an estimation rather than the ground truth. A possible improvement will be to obtain a different authoritative dataset that contains the official infection rate of each county or other rate that can be used as our ground truth. For example, 1point3acres has fatality rates of each county in the U.S. that can be a great candidate for this project's ground truth. Unfortunately, we never heard back from 1point3acres after several attempts to ask for data access. Besides, we are

aware that the number of positive COVID-19 cases largely depend on the number of COVID-19 tests conducted. For the counties that did not hold any COVID-19 tests or which did not report the cases, their infection rates are zeros. Inaccurate data leads to inaccurate predictions. Furthermore, the team picked a specific date, 04/12/2020, to focus our analysis and predictions on, which could lead to certain bias in our data. Counties that had been exposed to the virus long before 04/12/2020 may have a very different infection rate from the counties that were exposed to the virus later. Acknowledging this selection bias, we understand that our predictions could potentially be improved if we have the precise date of the first positive case found in each county and specify our target variable to be the infection rate 2-4 weeks after the virus was first found in the county or the infection rate at the peak of the COVID-19 pandemic. This would allow us to capture the spread of the virus in a specific period of time.

Last but not least, we only used aggregated U.S. census data. The U.S. census data does not include all possible factors that directly affect the spread of the virus. For example, we do not know how many flights entering and leaving each county has every day, which may be critical to spreading the virus as transportation plays an important role in prediction, analyzed by our models. Besides, public policies such as quarantine or shut down at midnight could also affect the spreading speed. They are not taken into consideration to build the models due to the fact that they are too difficult to capture in our models and quite subjective.

To understand the relations between each feature and the infection rate, we plot out the coefficients from the Logistic Regression. Different from the prediction model in which features related to commute are given more absolute weight, classification by Logistic Regression relies more on features related to income level, education and renting. As shown in figure 5, large number of 'Households lived in Renter-Occupied Housing Units' leads to high probability of being in the high-risk county, while large number of 'Householder lived in Owner-Occupied Housing Units' leads to high probability of being in the low-risk county. 'Public Transportation' has the highest correlation with the infection rate but is given a weight close to zero by Logistic Regression. This counter-intuitive fact might be due to feature correlations present in all 97 features that are feeded into the model. A correlation analysis shows that 'Public Transportation' is correlated with many features Logistic Regression has given high weight on, such as 'Householder lived in renter-Occupied Housing Units', '$75,000 Or More', and 'Bachelor's Degree'. Therefore, the feature 'Public Transportation' could probably be represented as a combination of other features and thus was given a close to 0 weight. This finding also inspires us to do feature selection before running Logistic Regression. However, given the same effort of hyperparameter tuning, model performance after feature selection is worse on the test set than the original model. Thus we decide to use the Logistic Regression model without feature selection.

Based on our results, we can see that the COVID-19 prediction is not an easy task, and it is influenced by many factors. We would proceed with our project in the following aspects. First of all, it has been more than one month since 04/12/2020, the latest infection data could be considered to train the models. Furthermore, as mentioned above, we could specify our target variable to be the infection rate 2-4 weeks after the first positive case found in the county or the infection rate at the peak of the COVID-19 pandemic to capture the prevalence of the virus in a specific time period. Secondly, directly relevant information such as weather, traffic in and out from the county, etc. can also be taken into consideration. Besides the prediction task, we may also address the analysis of the results. For example, to figure out if counties with infection rates exceeding 1% have something in common, or to estimate how quarantine slows down the virus spreading speed.

6. Appendix

```
['AGE15_19', 'AGE20_24', 'AGE25_29', 'AGE30_34', 'AGE55_59', 'AGE60_64',
 'TOTAL_MALE', 'TOTAL_FEMALE', 'PERCENT_FEMALE', 'PERCENT_MALE',
 'AGEUNDER15', 'AGE35_44', 'AGE45_54', 'AGE65_74', 'AGE_75OVER',
 'PERCENT_TOTALPOP', 'INCOME_PERCAPITA', 'WORKERS 16 YEARS AND OVER',
 'CAR, TRUCK, OR VAN', 'WORKERS PER CAR, TRUCK, OR VAN',
 'PUBLIC TRANSPORTATION (EXCLUDING TAXICAB)', 'WALKED', 'BICYCLE',
 'TAXICAB, MOTORCYCLE, OR OTHER MEANS', 'WORKED AT HOME',
 'WORKED OUTSIDE COUNTY OF RESIDENCE',
 'WORKED OUTSIDE STATE OF RESIDENCE',
 'WORKERS 16 YEARS AND OVER WHO DID NOT WORK AT HOME',
 'LESS THAN 10 MINUTES', '10 TO 14 MINUTES', '15 TO 19 MINUTES',
 '20 TO 24 MINUTES', '25 TO 29 MINUTES', '30 TO 34 MINUTES',
 '35 TO 44 MINUTES', '45 TO 59 MINUTES', '60 OR MORE MINUTES',
 'MEAN TRAVEL TIME TO WORK (MINUTES)', 'LESS THAN HIGH SCHOOL GRADUATE',
 'HIGH SCHOOL GRADUATE (INCLUDES EQUIVALENCY)',
 'SOME COLLEGE OR ASSOCIATE'S DEGREE', 'BACHELOR'S DEGREE',
 'GRADUATE OR PROFESSIONAL DEGREE', '$1 TO $9,999 OR LOSS',
 '$10,000 TO $14,999', '$15,000 TO $24,999', '$25,000 TO $34,999',
 '$35,000 TO $49,999', '$50,000 TO $64,999', '$65,000 TO $74,999',
 '$75,000 OR MORE', 'MEDIAN INCOME (DOLLARS)',
 'BELOW 100 PERCENT OF THE POVERTY LEVEL',
 '100 TO 149 PERCENT OF THE POVERTY LEVEL',
 'AT OR ABOVE 150 PERCENT OF THE POVERTY LEVEL',
 'HOUSEHOLDER LIVED IN OWNER-OCCUPIED HOUSING UNITS',
 'HOUSEHOLDER LIVED IN RENTER-OCCUPIED HOUSING UNITS',
 'LABOR_FORCE_PARTICIPATION_RATE', 'LAB_16-19', 'LAB_20-24', 'LAB_25-29',
 'LAB_30-34', 'LAB_35-44', 'LAB_45-54', 'LAB_55-59', 'LAB_60-64',
 'LAB_65-74', 'LAB_OVER75', 'EMPLOYMENT/POPULATION_RATIO', 'EMP_16-19',
 'EMP_20-24', 'EMP_25-29', 'EMP_30-34', 'EMP_35-44', 'EMP_45-54',
 'EMP_55-59', 'EMP_60-64', 'EMP_65-74', 'EMP_OVER75',
 'UNEMPLOYMENT_RATE', 'UNEMP_16-19', 'UNEMP_20-24', 'UNEMP_25-29',
 'UNEMP_30-34', 'UNEMP_35-44', 'UNEMP_45-54', 'UNEMP_55-59',
 'UNEMP_60-64', 'UNEMP_65-74', 'UNEMP_OVER75',
 'PERCENT WITH PRIVATE HEALTH INSURANCE',
 'PERCENT NO PRIVATE HEALTH INSURANCE',
 'PERCENT WITH PUBLIC HEALTH INSURANCE',
 'PERCENT NO PUBLIC HEALTH INSURANCE', 'PERCENT MISSING HEALTH',
 'LAND AREA(SQMI)', 'POP DENSITY'],
```

Figure 2: County Features

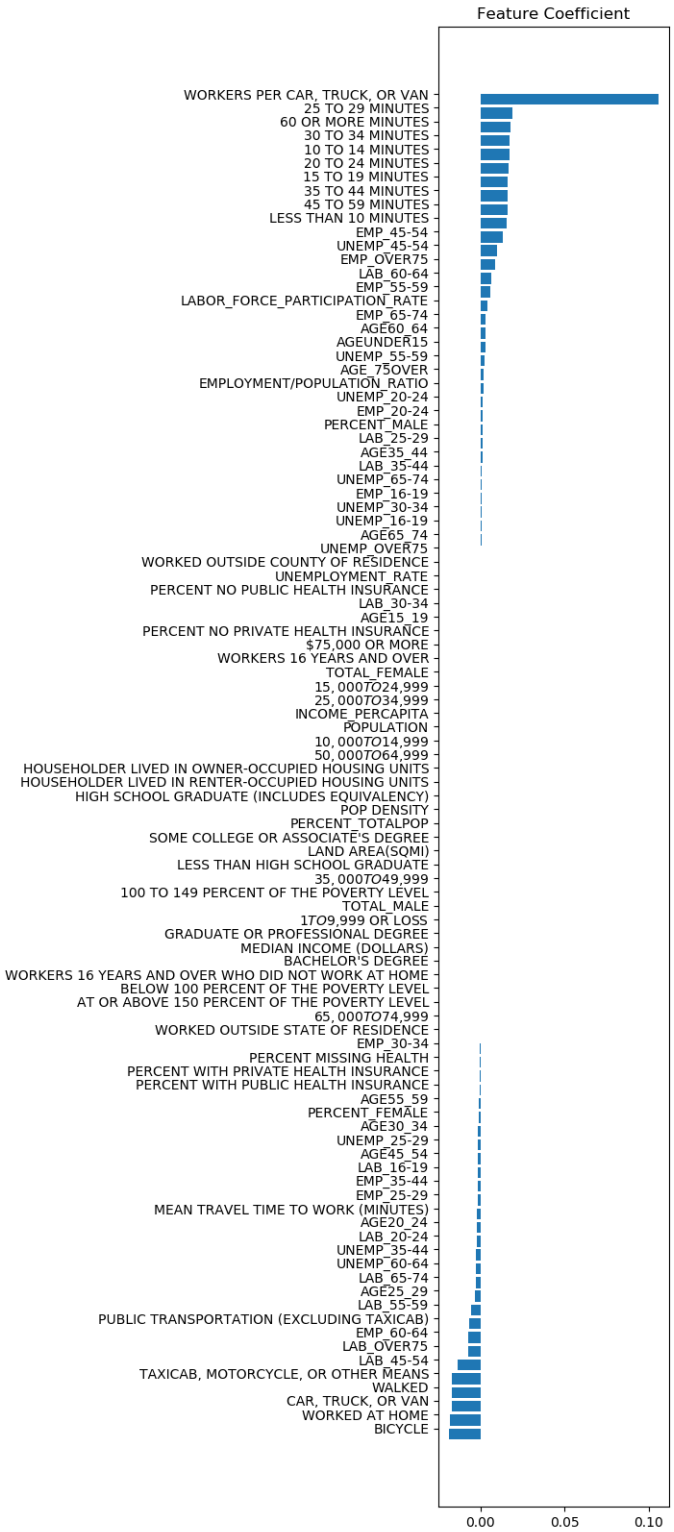

Figure 3: Feature Coefficients

```
['PUBLIC TRANSPORTATION (EXCLUDING TAXICAB)',
 'GRADUATE OR PROFESSIONAL DEGREE',
 "BACHELOR'S DEGREE",
 '$75,000 OR MORE',
 'LESS THAN 10 MINUTES',
 '$65,000 TO $74,999',
 'HIGH SCHOOL GRADUATE (INCLUDES EQUIVALENCY)',
 'AT OR ABOVE 150 PERCENT OF THE POVERTY LEVEL',
 '$25,000 TO $34,999',
 'AGE65_74',
 'TOTAL_FEMALE',
 'BELOW 100 PERCENT OF THE POVERTY LEVEL',
 'LESS THAN HIGH SCHOOL GRADUATE',
 'WORKERS 16 YEARS AND OVER',
 'WORKERS 16 YEARS AND OVER WHO DID NOT WORK AT HOME',
 '$1 TO $9,999 OR LOSS',
 '$15,000 TO $24,999',
 'LAB_16-19',
 'LAB_65-74',
 'PERCENT WITH PUBLIC HEALTH INSURANCE',
 'HOUSEHOLDER LIVED IN RENTER-OCCUPIED HOUSING UNITS',
 'UNEMP_20-24']
```

Figure 4: Selected Features

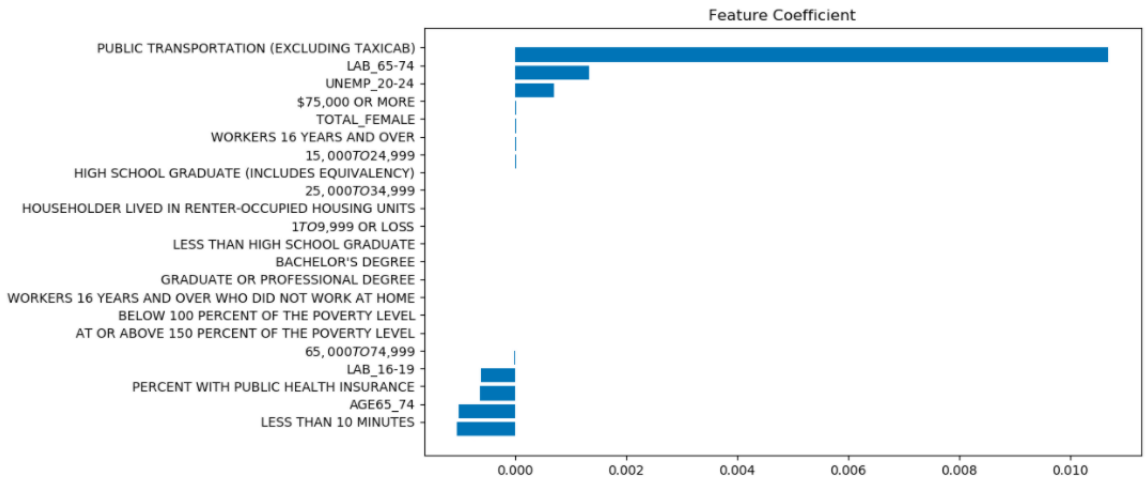

Figure 5: Selected Feature Coefficients

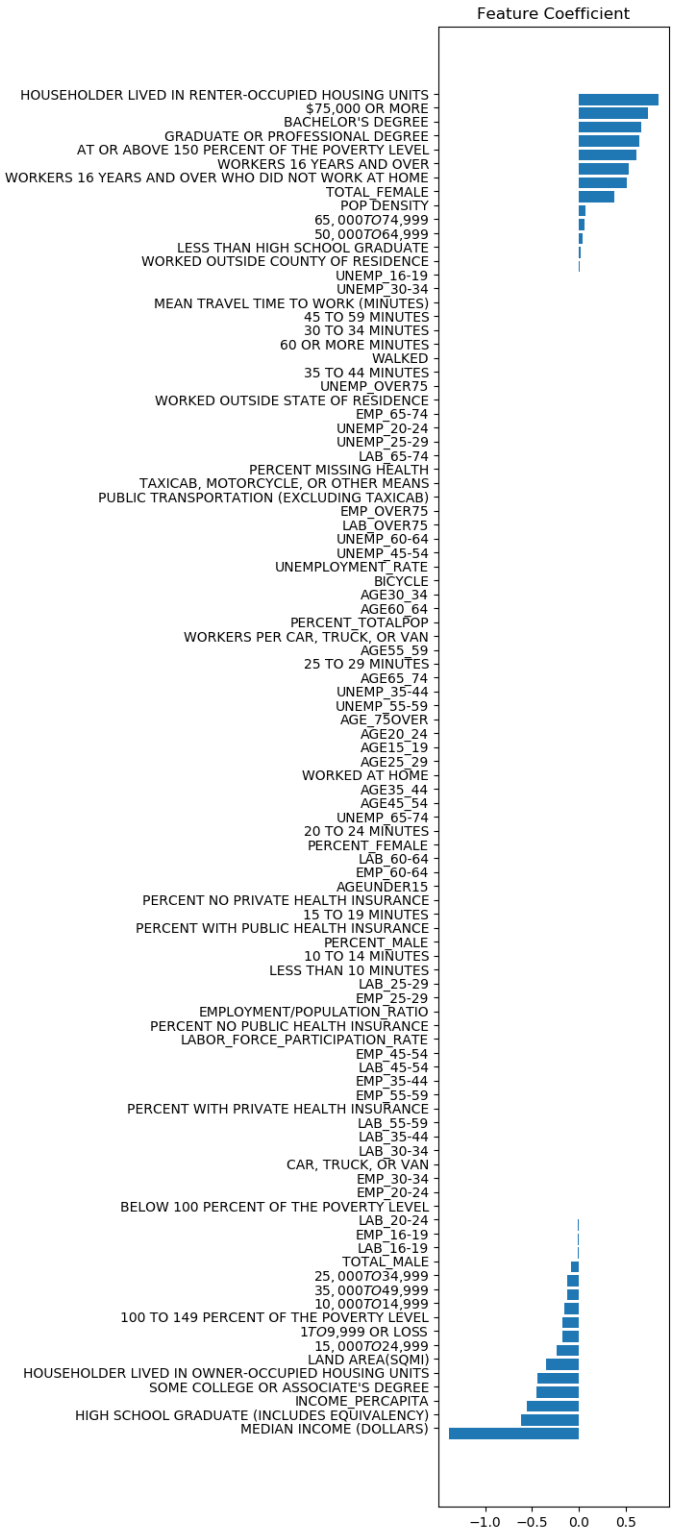

Figure 6: Feature Coefficients for Logistic Regression