CS229 Project Proposal

Courtney Noel Moran, Akwasi Sarpong Owusu-Akyaw, Richard Li

TOTAL POINTS

5 / 5

QUESTION 1

1 Proposal 5 / 5

✓ **+ 5 pts Complete**

+ 5 pts Complete, but we do recommend that you talk to a project TA.

+ 4 pts Missing key parts of motivation, methods or intended experiments. We recommend that you talk to a project TA.

Hello Richard, Courtney, Akwasi. Cool project idea! One suggestion is to frame the problem more as an ML task rather than as a statistical analysis / mathematical modeling project. For instance, it seems that 1-3 could be achieved without using any ML, and statistical analysis of correlations of different variables. That said, it sounds like the dataset can offer some interesting predictive tasks and that even if you go down the road of more stats/modeling, you can still focus on general ML techniques such as bias/variance analysis and feature analysis. - Leo

Disproportionate burden of pollution in California

Members:
Richard Li rli2014@stanford.edu
Courtney Moran moranc@stanford.edu
Akwasi Owusu-Akyaw akwasio@stanford.edu

Project Category: Application project

Motivation / Intended Experiments:
California has a wide range of infrastructure on which society relies – including highways, farms, wastewater treatment plants, and electricity generators. However, these infrastructure impose a pollution burden on local communities, and are disproportionately located in low-income regions.[1]

For our project, we would like to:

1) model the correlation between pollution exposure and income level in California,
2) identify the key features (i.e. types of pollution) that best explain the variation in income level (and other socioeconomic metrics), and
3) provide recommendations to policymakers to allocate public funding (i.e. CEC grants for electric vehicle deployment) to maximize benefit for impacted communities.

Dataset:
CalEPA has compiled a comprehensive dataset[2] that quantifies multiple sources of air and water pollution for every census tract in California (>8000). Included in each geographical region are indicators of health and socioeconomic vulnerabilities to pollution – such as poverty rate, incidence of asthma, and education level.

The dataset is both comprehensive and clean, which should allow us to focus our efforts on modeling, instead of data acquisition and cleaning.

[1] CalEnviroScreen report: https://oehha.ca.gov/media/downloads/calenviroscreen/report/ces3report.pdf
[2] CalEnviroScreen dataset: https://oehha.ca.gov/media/downloads/calenviroscreen/document/ces3results.xlsx

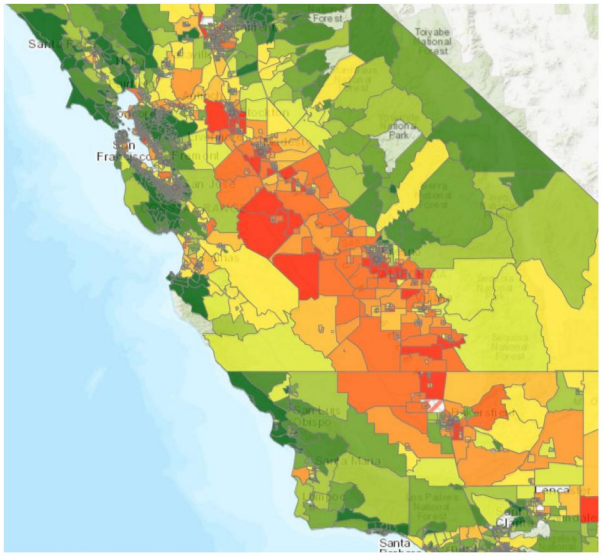

Fig 1: Pollution Index map of census tracts in California. Data provided by CalEPA.

Method:
Most of our features and target variables are continuous, so we will start with a simple linear regression model (scikit-learn). From there we can compare the performance of other models such as support vector machines. We are less concerned about developing a novel algorithm, as we are with choosing the right model to explain our data in a simple and robust way.

We would also like to evaluate the predictive strength of each of our features. By doing this, we can determine which factors of wealth show a strong correlation with our pollution data of interest. Sklearn provides different tools for feature selection that we can use.

Testing / Validation:
We will train our model using the majority of the data, and perform testing and validation on a withheld dataset. We can evaluate the performance of our model based on its accuracy in predicting labels in our validation dataset.

We will segment the examples by population density, so we can appropriately account for differences in metropolitan vs. rural regions. For instance, our city model trained on San Francisco could be used to predict pollution impacts in Los Angeles, but not in the Central Valley.

1 Proposal **5 / 5**

✓ **+ 5 pts Complete**

+ 5 pts Complete, but we do recommend that you talk to a project TA.

+ 4 pts Missing key parts of motivation, methods or intended experiments. We recommend that you talk to a project TA.

Hello Richard, Courtney, Akwasi. Cool project idea! One suggestion is to frame the problem more as an ML task rather than as a statistical analysis / mathematical modeling project. For instance, it seems that 1-3 could be achieved without using any ML, and statistical analysis of correlations of different variables. That said, it sounds like the dataset can offer some interesting predictive tasks and that even if you go down the road of more stats/modeling, you can still focus on general ML techniques such as bias/variance analysis and feature analysis. - Leo