# Carbon Monoxide Monitor Individual Report

Nick Harbeck

**1. Introduction.**

Carbon Monoxide or CO poisoning is a dangerous condition that can cause serious tissue damage, or even death.[1] Many U.S. households are at risk of CO poisoning due to the use of fuel-burning appliances inside the home. This project uncovers the trends and risk factors associated with CO poisoning in U.S. households. During our analysis, we found that household income, rent, educational attainment, household heat source, and the year the household was constructed were the most significant predictors of CO monitor status.

To date, very little research has been performed on the analysis of homes at risk of CO poisoning. The U.S. Consumer Product Safety Commission (CPSC) provides several resources on the risks of CO poisoning[2] and this report provides additional empirical analysis of the CO poisoning risk in U.S. households.

A CPSC memorandum on the risks of CO poisoning found that CO monitors reduce the risk of death in a CO poisoning incident.[3] Conversely, households without monitors are at a much higher risk of death from CO poisoning. This report helps to prioritize constituent groups most at risk of CO poisoning which could be used for targeted policies that reduce the risk of CO poisoning.

We specifically analyze 14 predictors of CO monitor status in a U.S. household (present or not present) from the American Housing Survey. A combination of classification algorithms was used to predict the presence of a CO monitor in a U.S. household. This report contains detailed information on the dataset, data mining process, experimental setup, results, and opportunities for future research.

**2. Description of Individual Work**

Most of the work that I did on this project was for the preprocessing of the data and the deployment of models. Because the dataset was so large, I subset the data into 18 columns that we wanted to analyze. This portion of the project was done in R because I found it easier to handle the datatypes as originally included the Census Bureau's .csv file. The .csv files of our data were also saved onto my website because of the large file size.

I also performed the data preprocessing and cleaning in python with the help of Ben. We tried to incorporate the data dictionary into the original dataset, but the complicated merging that would need

[1] Carbon monoxide poisoning. 2019. Mayo Clinic. Accessed from https://www.mayoclinic.org/diseases-conditions/carbon-monoxide/symptoms-causes/syc-20370642

[2] Carbon Monoxide Information Center. 2019. United States Consumer Product Safety Commission. Accessed from https://www.cpsc.gov/Safety-Education/Safety-Education-Centers/Carbon-Monoxide-Information-Center

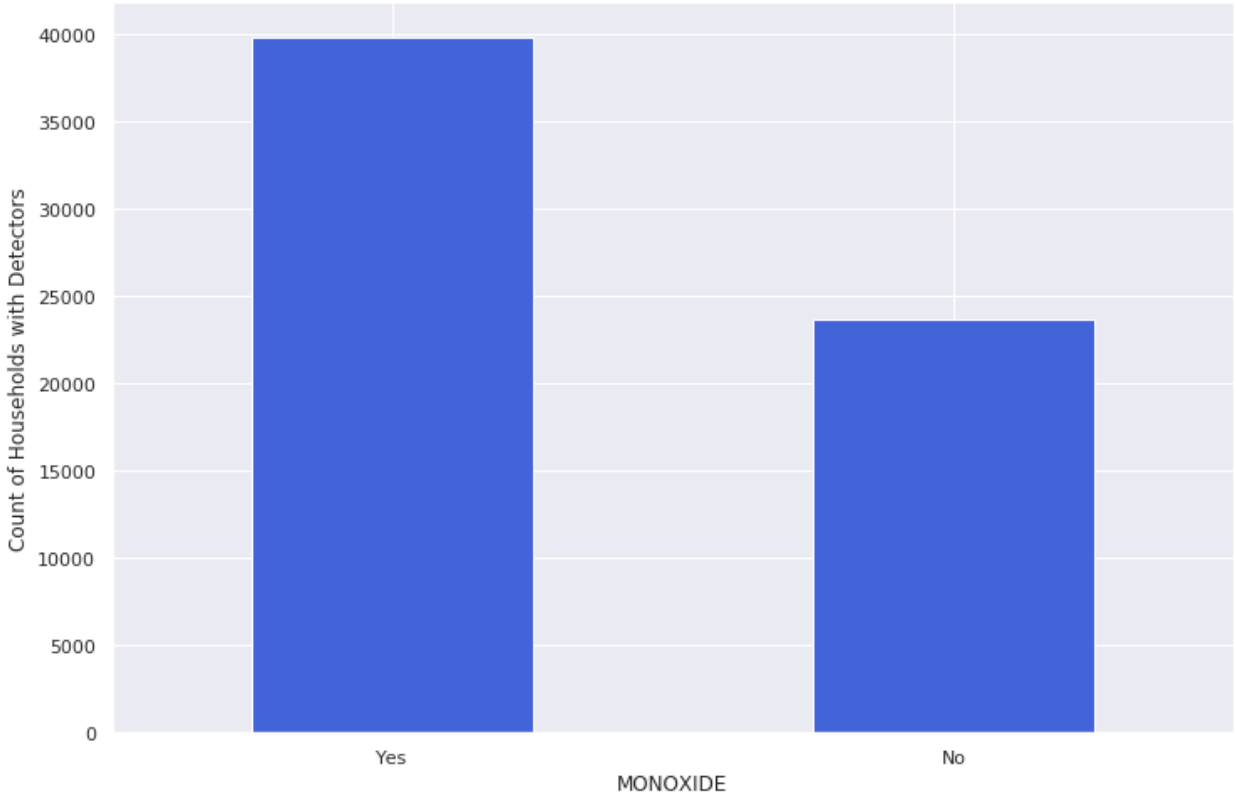
[3] Ault, Kimberly. Estimates of Non-fire Carbon Monoxide Poisoning Deaths and Injuries. 1997. Accessed from https://www.cpsc.gov/s3fs-public/pdfs/foia_3512c1f.pdf

to take place meant that we were better off interpreting the specific results that we needed for the project. Below is a copy of the code that processes the raw data:

```
for col in
COdf[['MONOXIDE','CONTROL','HEATTYPE','HEATFUEL','HOTWATER','ACPRIMARY'
,'ACSECNDRY','HOA','RENTCNTRL','HHMAR','HHGRAD','HHRACE','INTLANG']]:
  COdf[col] = COdf[col].apply(lambda x: re.findall('[^0-9-]+([0-9-]+)',
str(x)))
  COdf[col] = COdf[col].str.get(0)
  COdf[col] = pd.to_numeric(COdf[col])
```

To help understand the data, I added in some exploratory analysis of some general information about the dataset include feature response distributions.

**Most U.S. Households have a CO Monitor**




The code below also outputs response distributions for each of the features we analyzed:

```
def col_histogram(df,column):
    plt.figure(figsize=(12,8))
    sns.set(style="darkgrid")
    sns.distplot(df[column].values)
    plt.tight_layout()
    plt.show()
features =
['MONOXIDE','CONTROL','HEATTYPE','HEATFUEL','HOTWATER','ACPRIMARY','ACS
ECNDRY','HOA','RENTCNTRL','HHMAR','HHGRAD','HHRACE','INTLANG']
```

```
for i in features:
  col_histogram(df,i)
```

The next step was to run the three different classifiers; logistic regression, random forest, and support vector machine. We used these three classifiers because CO monitor status has a binary outcome (present or not present) and we wanted multiple models to increase the robustness of our results.

This project was also an opportunity for me to apply the pipeline from scikit-learn that is covered in Machine Learning 1. This tool is helpful because it simplifies the different steps that are needed to obtain a reliable model like scaling and cross validation.

```
def cvf(pipe,X,y, n_splits):
    accs = cross_val_score(pipe,
                           X,
                           y,
                           cv=KFold(n_splits=n_splits, random_state=0))
    print('The average accuracy score of the model is ', round(accs.
mean(), 3))
    print('The std deviation of the accuracy score is ', round(accs.
std(), 3))
    return round(accs.mean(), 3)
```

```
      pipe_logit = Pipeline([('StandardScaler', ss), ('Logistic', lo
git)])
pipe_rf = Pipeline([('StandardScaler',ss), ('Random Forest', RandomF
orestClassifier(n_estimators=100, max_depth=10, min_samples_leaf=10,
 random_state=0))])
pipe_svm = Pipeline([('StandardScaler',ss), ('SVC', SVC(random_state
=0))])
```

After running all three models, I added in an ensemble method using VotingChoice, but the model's accuracy did not increase.

```
model Vote ...
The average accuracy score of the model is  0.285
The std deviation of the accuracy score is  0.076
0:00:06.538192

 ------------------------------
accuracy  -  model with Kfolds=5 cross-validation
0.285 - Vote
```

Nevertheless, logistic regression and random forest models contain coefficients and feature importance which help with model interpretability. I needed to do this in two rounds because the first round that included all features did not have models with high accuracy.

## 3. Results

Logistic regression, random forest, and support vector machine classifiers were used to analyze the dataset. This is because our labels for CO monitor status are not present or present. Initial exploration of the data found that most households have a CO monitor, but a significant portion do not have one.

Our analysis found that certain features, including household income, rent status, year household was built, education level of the occupants, and the type of heating appliances are linked to the presence of a carbon monoxide monitor in U.S. household.

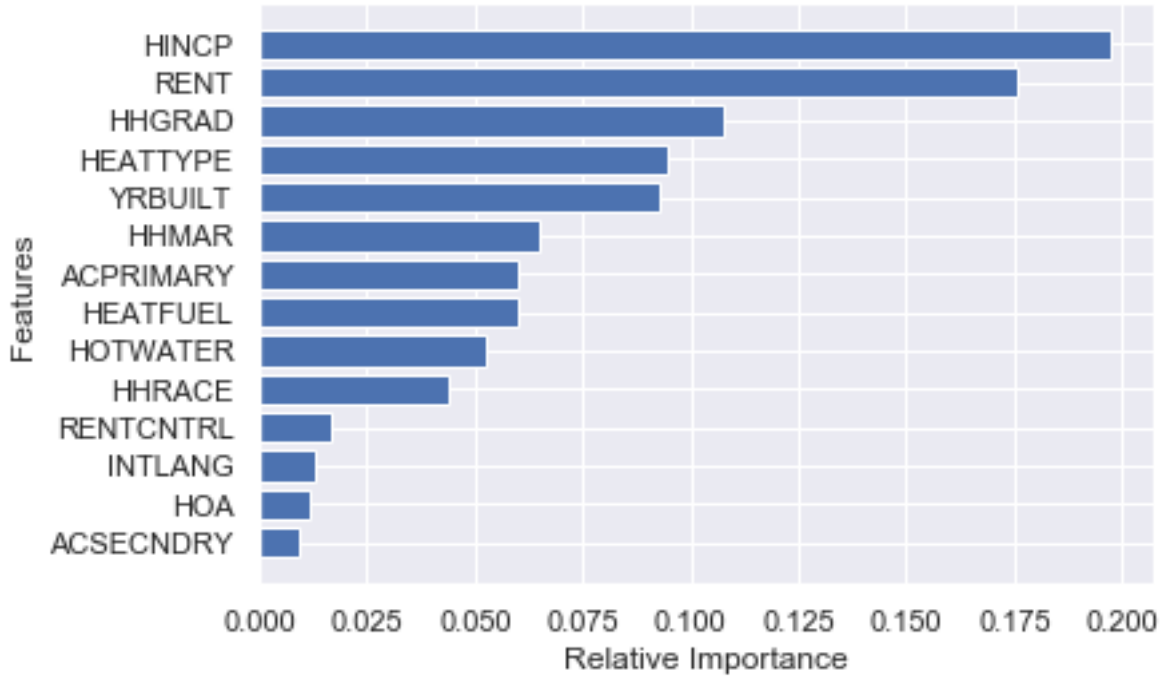
Python was the main tool to perform this analysis because it efficiently handles large datasets and a significant number of packages are available for data science work.

The cross validated and trained data models each obtained the following accuracies:

```
model Logistic ...
The average accuracy score of the model is  0.182
The std deviation of the accuracy score is  0.094
0:00:00.085932

 model Random Forest ...
The average accuracy score of the model is  0.401
The std deviation of the accuracy score is  0.057
0:00:01.671271

 model SVC ...
The average accuracy score of the model is  0.333
The std deviation of the accuracy score is  0.066
0:00:04.589659

 ------------------------------
accuracy  -  model with Kfolds=5 cross-validation
0.401 - Random Forest
0.333 - SVC
0.182 - Logistic
```

Below are feature importance for the dataset and coefficients to help predict CO monitor status.

**According to our model, the most important features are income, rent status, education level, heat type, and year built**




These accuracies were not good, so I ran the model again using only the five most important features

```
 model Logistic ...
The average accuracy score of the model is  0.566
The std deviation of the accuracy score is  0.066
0:00:00.335734

 model Random Forest ...
The average accuracy score of the model is  0.622
The std deviation of the accuracy score is  0.05
0:00:06.958918

 model SVC ...
The average accuracy score of the model is  0.614
The std deviation of the accuracy score is  0.043
0:02:26.778303

 ------------------------------
accuracy  -  model with Kfolds=5 cross-validation
0.622 - Random Forest
0.614 - SVC
0.566 - Logistic
```

The classification report for the most accurate model also had better results.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.67 | 0.63 | 0.65 | 10675 |
| 2 | 0.65 | 0.69 | 0.67 | 10639 |

| | | | | |
|---|---|---|---|---|
| accuracy | | | 0.66 | 21314 |
| macro avg | 0.66 | 0.66 | 0.66 | 21314 |
| weighted avg | 0.66 | 0.66 | 0.66 | 21314 |

In fact, when running the model with household income as the only predictor, we found that accuracy was approximately 58%.

For each of these features, we ran a logistic regression to obtain the coefficients of each response option. This allows us to see how the most important features, as indicated by the random forest model, affect the likelihood of CO monitor presence in a household.

We can see how the features change the likelihood of CO monitor presence in a household below:

| | |
|---|---|
| HOTWATER | -6.446435e-02 |
| ACSECNDRY | -4.420390e-02 |
| HEATFUEL | -3.032724e-02 |
| **HHGRAD** | **-1.845504e-02** |
| HOA | -3.016613e-03 |
| HHMAR | -8.478037e-04 |
| **RENT** | **-1.140057e-04** |
| **HINCP** | **5.251367e-07** |
| **YRBUILT** | **6.651862e-04** |
| INTLANG | 6.890235e-03 |
| ACPRIMARY | 1.730811e-02 |
| RENTCNTRL | 1.740987e-02 |
| HHRACE | 3.376590e-02 |
| **HEATTYPE** | **3.546891e-02** |

Higher income, gas-fueled appliances, lower rent, newer homes, and lower education levels increase the likelihood of a household with a CO monitor. Some of these results are interesting because we expected older homes, lower rent, and lower education households to be less likely to have a CO monitor because these features tend to be correlated with lower household income.[4] This is an area that would be a good direction for future analysis

**4. Summary and conclusions.**

CO monitor status in U.S. households can be predicted from a variety of factors listed in the 2017 AHS. Our results allow us to conclude that any efforts to increase the presence of CO monitors in U.S. households should target low income households first. This is because household income is the most important feature according to our random forest classifier and lower income households are less likely than higher income households to have a CO monitor according to the logistic regression model.

We also saw some counterintuitive results in the models that we ran. For example, lower rent, and lower education households all tend to increase the likelihood of a household having a CO monitor. The lower rent and graduate status could also be features that are closely correlated with households that have gas appliances. If there is no source of gas combustion (in the case of higher rent and higher

[4] Douglas-Hall, Ayana and Chau, Michelle. 2007. Parents' Low Education Leads to Low Income, Despite Full-Time Employment. Accessed from http://www.nccp.org/publications/pub_786.html

education status households), CO generation, a byproduct of gas combustion, is much less likely to occur in a home and the importance of having a CO monitor diminishes.

The 2017 AHS dataset offers many opportunities for data analysis and mining. In the future, it would be helpful to look at different models, optimize model hyperparameters, and expand the scope of features that can be used to predict the presence of a CO monitor in a U.S. household. Further research could also be performed to combine survey results with CO incidents in the U.S. The Federal Emergency Management Administration (FEMA) reports chemical release incident data for the U.S. and lists incidents according to chemical type and location.[5] The geographic indicators in the 2017 AHS could be combined with the FEMA data to look at how CO poisoning severity links to demographic and other appliance factors as listed in the AHS dataset.

**5. Percentage of Code from the Internet**

Approximately 9% of the code that I used (10/108 lines) came from the internet. However, this number is not indicative of the amount of original code that I created because I was able to copy a significant portion of model codes from the Machine Learning 1 class. Approximately, 44% (45/108 lines) of my code was modified from previous work.

[5] National Fire Incident Reporting System. Accessed from https://www.nfirs.fema.gov/