Ben Thompson
DATS 6103.17: Data Mining
Prof. Jafari

Individual Final Report

Project Description & Shared Work:

My group consisted of Nick Harbeck, Jacob King, and myself. Our project centered around using the 2017 AHS Survey data to predict whether surveyed households were at risk of CO exposure due to their lack of having a CO Monitor. To this end, our analysis used the AHS 2017 Survey's *MONOXIDE* feature (a binary variable indicating if respondents had a CO detector) as our target. We then leveraged a variety of the AHS survey's other demographic and environmental features to construct predictive models for CO Monitor ownership.

Ultimately, we were able to run several models (random forest, SVM, and logistic regression) and also use additional ensembling methods to obtain a prediction of CO monitor ownership with ~62% accuracy. According to our K-Fold crossvalidation, our analysis indicated that household income, rent status, household educational attainment, the year a home was build, and the type of heating used in a household were the most important predictors of CO monitor ownership. As we described in our report, higher incomes, gas -fueled appliances, lower rent, and older homes coincided with an increased likelihood of households owning a CO monitor.

In terms of shared work, each group member played a role in thinking about the models to be deployed in the analysis. Nick and myself also shared the responsibility of sourcing the raw AHS data, and narrowing it down into a subset of data that would be useful for our analysis. Nick took the lead on designing the logistic, SVM, and RF models – as well as piping them into a K-Folds cross-validation model. Nick also summarized these results and drafted the report on them. Jacob implemented an additional logistic regression ensembling method and summarized these results. Nick drafted the overall report and also put together a short powerpoint which Jacob and I reviewed.

Individual Work:

As mentioned earlier, part of my initial work involved sourcing AHS materials. This involved thinking about the kinds of variables that would be suitable for our analysis – as well as the nature of the AHS Survey data itself. I selected several of the demographic variables to include in the analysis. I also sourced the codebook for the AHS 2017 data (which in its raw form is uncoded). Unfortunately, I was not able to implement re-labeling via the codebook directly into our python dataset. This was due to difficulties with the formatting of the codebook versus the AHS data itself. However, the codebook did help in narrowing down the variables we would need to ultimately consider – and at this point, because we narrowed our data, the codebook's labels were no longer necessary to directly substitute into the AHS dataset that we analyzed.

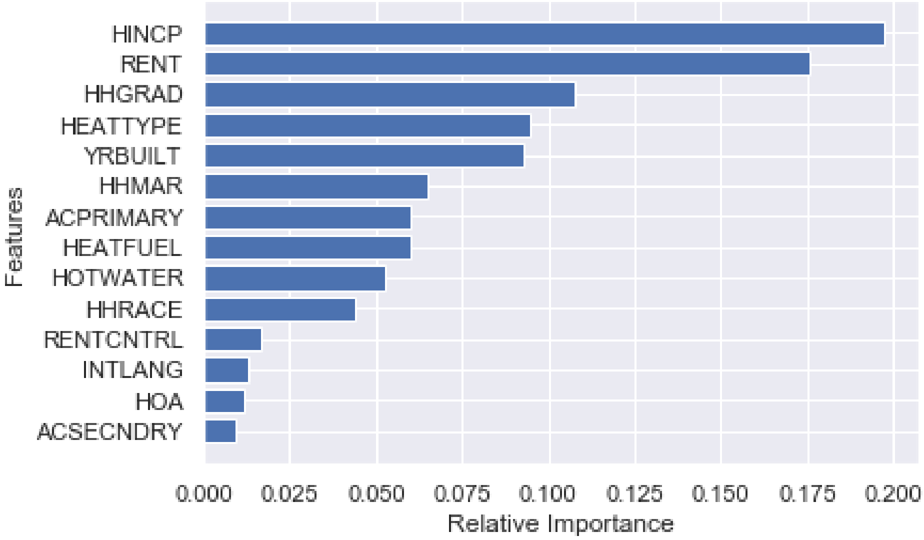
The other large portion of my work involved constructing the GUI for our presentation. I was unfamiliar with constructing GUIs prior to this assignment, and so I spent several hours evaluating and testing out several different GUI options. I initially started to construct a GUI using *TKinter*, however, its combability with *Matplotlib* seemed very limited. Using *Matplotlib* was essential to our analysis, and so I switched the intended GUI constructor over to *QT Designer* and *PyQT 15*.

While constructing the GUI itself was time consuming, I found learning how to use the GUI and gaining familiarity with its commands to be the most challenging – and I had to spend several hours researching how to use *PyQT 5* and how to deploy *Matplotlib* within it. This research time was extended by the fact that *PyQT 5* recently superseded *PyQT 4* and that most of the instructional materials currently available online still reference *PyQT 4*. Between the two versions, there are notable difference in how modules are packaged – meaning that I had to read a good deal of documentation in *PyQT 5* to figure out how to execute similar commands from *PyQT 4*.

It was also challenging to deploy *Matplotlib* within *PyQT 5*. While I did find a tutorial on how to deploy *Matplotlib* in a *PyQT 4* shell and was able to adjust it to *PyQT 5*, our project relies on specific uses of *Matplotlib* that I had to figure out how to deploy in a GUI.[1] Our GUI does this by using *Matplotlib*'s *pyplot.figure* module. This adds challenges because conventional *Matplotlib* commands often did not apply – and in addition, may of the *.figure* commands are not be flagged by python as immediately valid options. So, again, I had to read a lot of documentation on how to deploy this. Thankfully, I succeeded in getting most of our project's figures into the GUI – allowing users to more easily view the AHS data that went into our analysis, as well as the findings of our analysis.

Results:

In terms of our analysis's results, we found fairly consistent performance across each of the modeling approaches that we employed. Our logistic model attained a predictive accuracy on our test set of 56.6%. Our RF model also achieved 62.2% and our SVM model achieved 61.4%. The standard deviations of each model's accuracy were all the same within three decimals. We ensembled these individual models using K-Folds cross-validation and found that household income most important predictor of CO ownership – followed by rent status, household educational attainment, household heating method, and the year that the house was built.



[1] R.C.Nelson, "Building a Matplotlib GUI with Qt Designer", http://blog.rcnelson.com/building-a-matplotlib-gui-with-qt-designer-part-1/.

Applying Kfolds yields an accuracy of 62% on our test predictions. And ensembling via bagging results in a model vote at an accuracy of 28.5% again.

K-Folds Confusion Matrix/Accuracy Statistics:

```
                    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
```

Bagging Ensembling:

```
      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
```

Ensembling with logistic regression did not improve the accuracy of our model and so we defer to the random forest model.

Summary & Conclusions:

To summarize the results of our model across these coefficients in terms of logistic regression log-likelihood coefficients, we observed higher incomes, ownership of gas-fueled heating units, older homes, and lower levels of educational attainment all increase the likelihood that a household will own a CO monitor. This runs contrary to what we might have expected – since we would think that more educated household would be more knowledgeable about the dangers of CO. We were also surprised that lower rents, and older homes had increased likelihood for CO ownership. A more detailed analysis would be needed to see how these variable interact – as well as to explore what other factors might be impacting CO usage. Our models also did not factor in geographic areas – something we expect would impact households' choice of heating units, income, education, and other important factors that the current model identified.

It would be advantageous as well if we were able to run feature selection on a larger portion of the AHS dataset. In this case, we purposively selected features we hypothesized would be relevant – based on my knowledge of survey data and Nick's subject-matter expertise on HVAC. However, we might try to deploy a feature selection approach on the larger AHS and compare this approach's predictive capabilities with our own purposive selection.

We should also directly predict using our logistic model so that we can get a sense of the magnitude behind the probabilistic relationships that our report cites. Currently, it is limited to log-likelihood – which doesn't provide an immediately intuitive sense of magnitude for the relationships that we observe on CO ownership.

We might also spend more time on pre-processing in order to implement all codebook labels. This would increase the ease of interpretation as the analysis moves forward. In this case, we managed to carry out the analysis without a full labeling of the data – but it would be necessary to fully implement labels if we wanted to use a larger selection of the data later on – or if we wanted to experiment with more complex imputation strategies for missing data.

Code Calculation:

[NOTE: I am still writing more code for the GUI and so the % below will not reflect the full amount of unique code that I have written]

Copied: 30 lines
Modified: 15 lines
Added: 84 (not final)
Overall taken from internet: 17%