Matt Alvarez-Nissen and Lilla Petruska
MS&E 226
11/19/2020

MS&E 226 Project Part 2 - Neighborhood Outages

We investigated Pacific Gas & Electric (PG&E) power outages and demographic factors that influence their scope (temporally and spatially). Specifically, we asked the question: are there factors in census tracts that impact PG&E's service to its customers and response to such power outages?

Part 1: Prediction on the test set

Regression

Our best performing linear regression model transformed two covariates, prop_latino and prop_less_than_hs. We selected this model because it had the lowest cross-validation error (4.065). While still having a low R^2 and relatively high CVerror, this model is able to pick up on small changes in the prop_latino and prop_less_than_hs covariates, as they have heavy-tailed distributions, without overfitting the training data. Therefore, we believed it would be more generalizable to the test data.

```
lr_transform <- lm(median_outage_duration_hr ~ ., data = reg_train_transform)

transform_predict <- predict(lr_transform, reg_test_transform)
RMSE(reg_test_transform$median_outage_duration_hr, transform_predict)
```

```
## [1] 3.401094
```

The resulting RMSE is 3.40194, which is lower than the CVerror. The fact that our test data fit our model better than our training data may have occurred simply due to chance. However, we did select the model because it has the best generalization error and it seems to have stood up to that test.

Classification

Our best performing classification model used principal components analysis (PCA) to reduce variables and to account for the collinearity of our covariates. Following PCA, we then used K-Nearest Neighbors (KNN) to determine if an outage affected more than the average amount of customers. Our focus was on reducing 0-1 loss and increasing sensitivity.

On the test set, it is apparent that mean 0-1 loss increased significantly and sensitivity dropped. Other metrics improved, like precision, while others worsened, like the Type II Error Rate. Given our focus on 0-1 loss and sensitivity, this model had an underwhelming performance. It would have been preferable to maintain a higher sensitivity rate with a lower 0-1 loss, but it is not clear how possible this is due to the low explanatory nature of the covariates.

Part 2: Inference

Parts a) and b) below

```
# variable selection (non-PCA)
# start with basic logistic model (train)
logit_mod_train <-
  glm(
    above_median_cust_affected ~ .,
    data = acs_outages_class_train,
```

Table 1: Above Median Customer Affected Confusion Matrix Metrics

metric	train	test
Mean 0-1 Loss	0.0088275	0.0364862
Precision	0.6307692	0.8333333
Sensitivity	0.0953488	0.0454545
Specificity	0.9861512	0.9976690
Type I Error Rate	0.0138488	0.0023310
Type II Error Rate	0.9046512	0.9545455
False Discovery Rate	0.3692308	0.1666667

```
    family = binomial
  ) %>%
  # stepwise selection by AIC
  MASS::stepAIC(trace = FALSE, direction = "both")
summary(logit_mod_train)
```

```
##
## Call:
## glm(formula = above_median_cust_affected ~ prop_white + prop_black_or_african_american +
##     prop_american_indian_and_alaska_native + prop_asian + prop_multi_racial +
##     prop_owner + prop_rural + n_outages_sq_km, family = binomial,
##     data = acs_outages_class_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6026  -0.7078  -0.5210  -0.3080   2.6176
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.862301   0.285359 -10.031  < 2e-16
## prop_white                           3.331315   0.366626   9.086  < 2e-16
## prop_black_or_african_american       2.755705   1.098773   2.508   0.0121
## prop_american_indian_and_alaska_native  9.707541   4.096658   2.370   0.0178
## prop_asian                           2.863358   0.432034   6.628 3.41e-11
## prop_multi_racial                   -6.759844   2.774238  -2.437   0.0148
## prop_owner                          -0.596982   0.313132  -1.906   0.0566
## prop_rural                           0.510446   0.212239   2.405   0.0162
## n_outages_sq_km                     -0.030591   0.006589  -4.642 3.44e-06
##
## (Intercept)                         ***
## prop_white                          ***
## prop_black_or_african_american      *
```

```
## prop_american_indian_and_alaska_native *
## prop_asian                             ***
## prop_multi_racial                      *
## prop_owner                             .
## prop_rural                             *
## n_outages_sq_km                        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2157.5  on 2162  degrees of freedom
## Residual deviance: 1962.0  on 2154  degrees of freedom
## AIC: 1980
##
## Number of Fisher Scoring iterations: 5
```

```
# extract just the selected variables to run through the bootstrap
acs_outage_var_select <-
  acs_outages_class_train %>%
  select(above_median_cust_affected, any_of(names(logit_mod_train$coefficients)))
```

For this task, we decided to use a logistic model for classification. Due to the high number of covariates, we used stepwise regression (both directions) to select for variables. The stepwise selection highlighted proportions of white, Black/African American, American Indian and Alaska Native, and multi-racial people, as well as proportion of home owners, proportion of rural land, and the number of outages per square kilometer (outage density). By definition, each chosen covariate in the final model is significant at some level. In other words, these variables are only significant in so far are they are significant compared to the excluded covariates. Therefore, we do not put a lot of stock in the results - as it's quite likely that there are other covariates with much stronger explanatory power that were not included or available in original dataset. Additionally, stepwise selection does not consider all possible models and may be missing important context. We ultimately chose to use stepwise selection due to computational limitations, the low explanatory power of the data itself, and the presence of collinearity. Based on these results, the proportion of Latino, Native Hawaiian or Pacific Islander, or some other race residents do not have a significant effect on outage extent. Nor does income status or educational attainment (none of the covariates related to these indicators are sufficient). This may mean that PG&E's service quality does not vary according to the socioeconomic makeup of a census tract, which is understandable as public utilities have a responsibility to provide their customers with reliable service. The proportion rural or urban land use is also significant, which is surprising as we might expect census tracts with more urban land use (meaning more homes, industrial buildings, etc.) and higher population density to have more intense outages.

```
# variable selection (non-PCA)
# start with basic logistic model (test)
logit_mod_test <-
  glm(
    above_median_cust_affected ~ .,
    data = acs_outages_class_test,
    family = binomial
  ) %>%
  # stepwise selection by AIC
  MASS::stepAIC(trace = FALSE, direction = "both")
summary(logit_mod_test)
```

```
##
```

```
## Call:
## glm(formula = above_median_cust_affected ~ prop_white + prop_black_or_african_american +
##     prop_asian + prop_multi_racial + prop_college + prop_high_school +
##     n_outages_sq_km, family = binomial, data = acs_outages_class_test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3065  -0.7403  -0.5222  -0.3093   2.5080
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.43500    1.21036  -0.359 0.719299
## prop_white                      4.75128    1.20415   3.946 7.96e-05 ***
## prop_black_or_african_american  5.28621    2.18169   2.423 0.015393 *
## prop_asian                      3.78158    1.21319   3.117 0.001827 **
## prop_multi_racial              12.19522    5.21282   2.339 0.019311 *
## prop_college                   -4.92332    1.97673  -2.491 0.012751 *
## prop_high_school               -4.37084    2.22846  -1.961 0.049836 *
## n_outages_sq_km                -0.05210    0.01451  -3.590 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 545.48  on 538  degrees of freedom
## Residual deviance: 500.16  on 531  degrees of freedom
## AIC: 516.16
##
## Number of Fisher Scoring iterations: 5
```

Running the logistic model and stepwise selection on the test data produced different results - proportions of white, Black/African American, and multi-racial people, as well as number of outages by square kilometer remain, but the test data suggests using proportion of Asian people, proportion of college-educated people, and proportion of high school-educated people. These differences are most likely due to the smaller sample size of the test set and the presence of bias. It is likely that in the census tracts available in the test set, it just so happens that education has a more significant relationship with outage extent than in the training data.

Part c)

```r
# bootstrap CI for each reg coefficient
# create function to return coefficients
boot_coefficients <- function(data, indices) {
  data <- data[indices,]

  logit_mod <-
    glm(
      above_median_cust_affected ~ .,
      # use reduced variables dataset
      data = data,
      family = binomial
    )
  coefficients(logit_mod)
}
```

Table 2: Coefficient Confidence Intervals

coefficient	value	conf	lower_ci	upper_ci
(Intercept)	-2.8623010	0.95	-3.4190719	-2.2678497
prop_white	3.3313148	0.95	2.5703322	4.0391978
prop_black_or_african_american	2.7557046	0.95	0.6125922	5.0022009
prop_american_indian_and_alaska_native	9.7075413	0.95	1.6037653	17.4131145
prop_asian	2.8633579	0.95	1.9866123	3.7056749
prop_multi_racial	-6.7598438	0.95	-12.5368716	-1.0409672
prop_owner	-0.5969823	0.95	-1.2366027	0.0406757
prop_rural	0.5104457	0.95	0.1074885	0.9377008
n_outages_sq_km	-0.0305912	0.95	-0.0454560	-0.0141785

```
# Need to adjust R value (number of replicates, not sure what to use)
logit_boot <- boot(acs_outage_var_select, boot_coefficients, R = 2000)

# Loop through the results and create table of coefficient CIs
coef_ci_table <- c()
for (i in 1:ncol(acs_outage_var_select)) {
  # run the boot CI on each coefficient
  result <- boot.ci(logit_boot, index = i, type = "norm")
  coef_ci_table <-
    result$normal %>%
    as.data.frame() %>%
    rename(lower_ci = V2, upper_ci = V3) %>%
    rownames_to_column(var = "coefficient") %>%
    bind_rows(coef_ci_table)
}

# merge CI table with original coefficient values
coef_comparison_table <-
  coefficients(logit_mod_train) %>%
  as.data.frame() %>%
  rownames_to_column(var = "coefficient") %>%
  rename(value = ".") %>%
  left_join(coef_ci_table, by = "coefficient")
coef_comparison_table %>%
  kable(caption = "Coefficient Confidence Intervals") %>%
  kableExtra::kable_classic() %>%
  #format for markdown
  kable_styling(latex_options="scale_down")
```

The results are the same as our standard regression output. No coefficients lie outside the range of the 95% confidence interval after running the reduced logistic model through the bootstrap. Some of the intervals of relatively large, however, which is of concern (i.e, for proportion multi-racial or proportion American Indian and Alaska Native). Additionally, the fact that proportion of homeowners has a confidence interval that changes signs is worth noting, and this is the only coefficient that was not significant below the 0.5 level whcih is consistent with our previous result.

Inference Discussion

```
# stepwise logistic model (train)
summary(logit_mod_train)
```

```
##
## Call:
## glm(formula = above_median_cust_affected ~ prop_white + prop_black_or_african_american +
##     prop_american_indian_and_alaska_native + prop_asian + prop_multi_racial +
##     prop_owner + prop_rural + n_outages_sq_km, family = binomial,
##     data = acs_outages_class_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6026  -0.7078  -0.5210  -0.3080   2.6176
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.862301   0.285359 -10.031  < 2e-16
## prop_white                           3.331315   0.366626   9.086  < 2e-16
## prop_black_or_african_american       2.755705   1.098773   2.508   0.0121
## prop_american_indian_and_alaska_native  9.707541   4.096658   2.370   0.0178
## prop_asian                           2.863358   0.432034   6.628 3.41e-11
## prop_multi_racial                   -6.759844   2.774238  -2.437   0.0148
## prop_owner                          -0.596982   0.313132  -1.906   0.0566
## prop_rural                           0.510446   0.212239   2.405   0.0162
## n_outages_sq_km                     -0.030591   0.006589  -4.642 3.44e-06
##
## (Intercept)                         ***
## prop_white                          ***
## prop_black_or_african_american      *
## prop_american_indian_and_alaska_native *
## prop_asian                          ***
## prop_multi_racial                   *
## prop_owner                          .
## prop_rural                          *
## n_outages_sq_km                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2157.5  on 2162  degrees of freedom
## Residual deviance: 1962.0  on 2154  degrees of freedom
## AIC: 1980
##
## Number of Fisher Scoring iterations: 5
```

```
# baseline logistic model (train)
# start with basic logistic model (train)
logit_base_train <-
  glm(
    above_median_cust_affected ~ .,
```

```
    data = acs_outages_class_train,
    family = binomial
  )
summary(logit_base_train)
```

```
##
## Call:
## glm(formula = above_median_cust_affected ~ ., family = binomial,
##     data = acs_outages_class_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4979  -0.7019  -0.5208  -0.3062   2.6394
##
## Coefficients: (5 not defined because of singularities)
##                                                Estimate Std. Error z value
## (Intercept)                                  -2.553e+00  1.342e+00  -1.903
## prop_white                                    3.038e+00  6.892e-01   4.408
## prop_black_or_african_american                2.662e+00  1.281e+00   2.079
## prop_american_indian_and_alaska_native        9.435e+00  4.340e+00   2.174
## prop_asian                                    2.537e+00  6.821e-01   3.719
## prop_native_hawaiian_and_other_pacific_islander  1.258e+00  6.707e+00   0.188
## prop_some_other_race                          3.717e+00  9.133e+00   0.407
## prop_multi_racial                            -6.932e+00  2.892e+00  -2.397
## prop_latino                                         NA         NA      NA
## pop_density_sq_km                             3.334e-05  2.255e-05   1.478
## prop_eli                                     -2.342e+00  2.723e+00  -0.860
## prop_hi                                      -7.690e-01  1.289e+00  -0.597
## prop_li                                      -1.453e+00  2.051e+00  -0.708
## prop_mi                                       2.409e-01  2.032e+00   0.119
## prop_vli                                            NA         NA      NA
## prop_college                                  4.837e-01  1.147e+00   0.422
## prop_high_school                              2.329e-01  1.210e+00   0.193
## prop_less_than_hs                                   NA         NA      NA
## prop_owner                                   -3.270e-01  3.976e-01  -0.822
## prop_renter                                         NA         NA      NA
## rental_vacancy_rate                           1.163e+00  1.093e+00   1.064
## owner_vacancy_rate                            2.839e+00  2.606e+00   1.089
## prop_rural                                    4.940e-01  2.266e-01   2.180
## prop_urban                                          NA         NA      NA
## n_outages_sq_km                              -3.509e-02  7.773e-03  -4.515
##                                              Pr(>|z|)
## (Intercept)                                    0.0571 .
## prop_white                                  1.04e-05 ***
## prop_black_or_african_american                 0.0377 *
## prop_american_indian_and_alaska_native         0.0297 *
## prop_asian                                     0.0002 ***
## prop_native_hawaiian_and_other_pacific_islander   0.8512
## prop_some_other_race                           0.6840
## prop_multi_racial                              0.0165 *
## prop_latino                                       NA
## pop_density_sq_km                              0.1393
## prop_eli                                       0.3898
## prop_hi                                        0.5507
```

```
## prop_li                                              0.4788
## prop_mi                                              0.9056
## prop_vli                                                 NA
## prop_college                                         0.6732
## prop_high_school                                     0.8473
## prop_less_than_hs                                        NA
## prop_owner                                           0.4108
## prop_renter                                              NA
## rental_vacancy_rate                                  0.2874
## owner_vacancy_rate                                   0.2759
## prop_rural                                           0.0292 *
## prop_urban                                               NA
## n_outages_sq_km                                   6.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2157.5  on 2162  degrees of freedom
## Residual deviance: 1956.2  on 2143  degrees of freedom
## AIC: 1996.2
##
## Number of Fisher Scoring iterations: 5
```

There is some change in significance of coefficients between the chosen model (which does not have all variables) and the full logistic model. For instance, the proportion of owners is not significant at all in the model that includes all covariates. Additionally, the proportion of Asian residents is highly significant in the full model, but is not present at all in the reduce model. In the model with all covariates, the proportion of rural land use is also significant and associated with a larger outages extent. Finally, the degree of significance between covariates in both models differs.

```
covariate_corr <-
  acs_outages %>%
  select(all_of(continuous_vars), all_of(outcome_vars)) %>%
  # drop prop owner
  select(-prop_owner) %>%
  select(where(~sd(., na.rm = TRUE) > 0)) %>%
  correlate() %>%
  stretch() %>%
  arrange(r) %>%
  filter(abs(r) > .5)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
# reformat in formula form
interaction_vars <-
  covariate_corr %>%
  mutate(lead_y = lead(y)) %>%
  filter(x != lead_y) %>%
  select(x, y) %>%
  unite(col = "interact", sep = ":") %>%
  unlist() %>%
```

```
  paste(collapse = " + ")
covariate_corr %>%
  group_by(r) %>%
  slice_head()
```

```
## # A tibble: 15 x 3
## # Groups:   r [15]
##    x                 y                      r
##    <chr>             <chr>              <dbl>
##  1 prop_rural        prop_urban        -1
##  2 prop_college      prop_less_than_hs -0.791
##  3 prop_college      prop_high_school  -0.786
##  4 prop_latino       prop_college      -0.762
##  5 prop_hi           prop_vli          -0.737
##  6 prop_hi           prop_li           -0.718
##  7 prop_hi           prop_mi           -0.675
##  8 prop_white        prop_less_than_hs -0.674
##  9 prop_white        prop_latino       -0.666
## 10 prop_hi           prop_renter       -0.629
## 11 prop_mi           prop_college      -0.569
## 12 prop_mi           prop_less_than_hs  0.504
## 13 prop_vli          prop_renter        0.530
## 14 pop_density_sq_km n_outages_sq_km    0.692
## 15 prop_latino       prop_less_than_hs  0.871
```

Collinearity is almost certainly impacting the results. When looking at the table above, we can see that all covariate pairs have an absolute correlation coefficient above 0.5. Many of these demographic variables (race/ethnicity, income level, education attainment) are related to one another. This is a fault with our project design and should have been considered as we built our dataset. The regression automatically attempts to account for this by removing one proportion from each demographic category. Even in the reduced model, there are multiple covariates that relate to proportions by race - which are certainly collinear variables. The most important impact of this is likely born out in the stepwise selection - the algorithm is choosing artificially inflated covariates to be part of the reduced model.

Using the stepwise approach during the model-building process likely biased the initial model in its determination of which covariates were most influential. Again, the issue of collinearity plagues our dataset, and the fact that so many of the final variables in our reduced model with racial demographic proportions highlights this fact. This reduction could have excluded important explanatory variables. In the future, we should consider external validation with new data and specifically integrating data from other California electric utilites and census tracts to strengthen the case for any particular associations.

We are not willing to determine any causal relationships or even associative relationships found in our inference, regardless of any significance determined by the regression for specific covariates. This is due to the persistant issues of collinearity and data completeness that we have discussed.

Part 3: Discussion

While PG&E and other electric utilities monitor their outages and service quality through numerous quantitative metrics, including spatial distribution, there is little emphasis on the communities that are affected by these disruptions in service. Of course, power outages impact residential, commerical, and industrial customers in different ways. However, we must also consider the equity implications of who is being impacted by irregular electricity service. Our models aimed to bring disparities in service quality to light by looking at the relationship between where outages are located and who they impact. Further, they try to help us understand if the demographic make up of a census tract makes a certain area more vulnerable to worse service quality or higher outage durations. Practically, this concept could be used in two spheres.

The first is disaster relief and preparation. In this case, the model would primarily be used for inference to determine what characteristics of census tracts may make an area more vulnerable, and thus highlight areas with similar makeups that could benefit from future investment in energy infrastructure or measures to prepare against large outages. When used for inference, the model could guide decisions in policy making related to resource allocation. The second sphere, in which the model would primarily be used for prediction, similarly would aim to highlight vulnerability but instead by predicting how long outages in a given area may be or how many residents could be impacted. This information would inform decision making for residents who may take action to prepare themselves for the event of particularly long outages. It is important to discuss here, however, that demographic data is likely not the best way to determine these insights. PG&E and other electric utilites making decisions about where and for how long to cut power must depend on all kinds of environmental and climate data, such as the presence of wildfires, temperature, and the quality of their transmission infrastructure, and thus we need to integrate these factors into our models in order to ensure more robust prediction and inference to better inform decision making.

Our models were fitted on power outage data from the last several months, but PG&E and other utilites are constantly collecting new data on their service regularity and variability. Patterns in the frequency and duration of power outages will likely change over time due to the prevalence of wildfires, improvements in transmission and distribution infrastructure, increased ability to predict wildfire behavior and movement, as well as changes in temperature and precipitation in California (which influence wildfire initiation). While the United States Census Bureau performs the census every ten years, the American Community Survey releases new data yearly. Our demographic data is thus relatively up to date and provides a current snapshot of census tract demographics. If we update our data every year with new demographic data, then we would expect better results from our models. However, this will impact our ability to use our models for inference because the covariates will be replaced each year. Therefore, our models need to be built to consider anticipated changes in both demographic and outage data while also considering historical data for inference. If possible, it would be best to give more weight to recent outage data when refitting because the newest data will best reflect contemporary trends.

We would want to make sure anyone working with our data or models know the following: When preparing our data, we calculated proportions for all our covariates and thus had to remove one covariate from each category (race/ethnicity, income level, education attainment, etc.) to avoid issues with collinearity. We designed our data to summarize statistics and outages at the scale of each census tract, which required making assumptions about the homogeneity of each tract. We believed this was reasonable, as census tracts are relatively small geographic areas. Correlations between covariates and their distributions were also considered when transforming data for our prediction model. For inference, we use stepwise selection to determine significant variables, which biases our selection of covariates and underestimates our p-values. We recommend validating our models on new data to mitigate this.

There are many things we should change regarding data collection. PG&E only provides a single coordinate for each outage. Therefore there is not reliable information concerning the extent of the entire affected outage area. We georeferenced the outage coordinate to determine which census tract it corresponded to, though this only allowed us to infer what communities a specific power outage affected. PG&E also does not provide a data dictionary, so we must infer what the different variables mean based on their names. Because we only used PG&E outage data, our model was design to elucidate utility service quality in primarily Northern California census tracts. Our entire dataset was only 2722 rows, which is relatively small and may have impacted our ability to find meaningful insights. There are other major electric utilites in California, such as Southern California Edison and San Diego Gas and Electric, that also serve geographies severely impacted by wildfires and with socioeconomic disparities. By building a dataset that included outage data on all 8,057 California census tracts, we would have a more complete picture of electric utility service in California. This would also allow us to understand differences in service between utilities.

When embarking on this topic, we believed that demographic characteristics of census tracts would have some predictive power to understand why PG&E's power outages were distributed the way they are. Historically, the quality of public services differ vastly based on communities' socioeconomic status, race, and wealth. After completing our analysis, there are likely many other factors at play in determining PG&E's ability to return power to its customers. Environmental and geographic factors, as well as PG&E's own behind-the-scenes decision making about where to cut power, will surely influence the prevalence and duration of power outages. Environmental and geographic factors that could have benefitted our model include the

size, proximity, and presence of wildfire; quality of transmission and distribution infrastructure (including powerlines and voltage boxes); more detailed land cover data (vegetation, temperative, albedo, forested area); and building infrastructure data (industrial and commercial offtakers use more electricity than residential customers). Perhaps it would have been wiser to try to model how PG&E itself makes choices about where and how long to cut power for.

When looking at our results, on one hand, it is a good thing that PG&E's service does not seemingly discriminate against certain groups based on the proportion of that demographic indicator present in a census tract. However, we believe that our research questions and the potential applications of this model are important things to think about and prepare for as wildfires and other natural disasters become more prevalent.