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

Matt Alvarez-Nissen and Lilla Petruska
MS&E 226
11/16/2020

MS&E 226 Project Part 2 - Neighborhood Outages

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.

Classification

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.

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,
    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)))
```

Ignore the following two chunks

```
# # start with basic logistic model (train)
# logit_mod_train <-
#   glm(
#     above_median_cust_affected ~ .,
#     data = acs_outages_log_train,
#     family = binomial
#   )
# summary(logit_mod_train)

# PCA logistic model (train) - NOT SURE IF WE SHOULD USE
# pca_logit_mod_train <-
#   glm(
#     above_median_cust_affected ~ .,
#     data = pca_table_train,
#     family = binomial
#   )
# summary(pca_logit_mod_train)
```

```
# # start with basic logistic model (test)
# logit_mod_test <-
#   glm(
#     above_median_cust_affected ~ .,
#     data = acs_outages_log_test,
#     family = binomial
#   )
# summary(logit_mod_test)

# PCA logistic model (test) - NOT SURE IF WE SHOULD USE
# pca_logit_mod_test <-
#   glm(
#     above_median_cust_affected ~ .,
#     data = pca_table_test,
#     family = binomial
#   )
# summary(pca_logit_mod_train)
```

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)
}

# 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")
```

Inference Discussion

Do you believe collinearity is impacting your results? Be specific.

Comment on post-selection inference: What aspects of your model-building and inference process that might bias your determination of which coefficients are significant?

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

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 a particular relationship.

Are confidence intervals wider or narrower on test set? THey should be wider because we are using less data.

Discussion

a) How would you expect your models to be used practically? Do you think they would primarily be used for prediction, for inference, or for both? What decisions do you think your models would guide, and what pitfalls do you see in using your models to make these decisions?

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. Suppose our model showed that certain covariates

PG&E and other electric utilites making decisions about where and for how long to cut power

b) How well would your models hold up over time (i.e., how often do you think they should be refitted)? Why?

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, and increased ability to predict wildfire behavior and movement. 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. Therefore, our model will need to be refitted regularly with the newest data on both outages and demographic data.

** how to integrate changes in land use, satellite data

c) Are there choices you made in your data analysis, that you would want to make sure any one (e.g., a manager, a client, etc.) that uses your models is aware of? Examples here might include approaches to data cleaning; data transformations that you chose; vulnerability to overfitting, multiple hypothesis testing, or post-selection inference; etc.

Our

d) If you could, how would you change the data collection process? In particular, are there reasonable covariates you would like to collect, that were not present in the data?

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 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.

e) If you were to attack the same dataset again, what would you do differently?

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.

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.

External validation using subsequent studies.