Homework 9

Add Your Name Here

Instructions

1. Add your name between the quotations marks on the author line in the YAML above.
2. Work through the problems below, composing your answer to each question between the bars of red stars.
3. Be sure to make commits often and push your commits to Github periodically.
4. When finished, knit your document to .pdf and view the resulting output to ensure it matches your intent.
5. To submit your assignment, make one final commit. Then make push your repo with all commits back to Github one final time.

Classification Competition

Your objective is to build a classification model **using the tools we have discussed in class** that predicts whether an email is spam, based on other characteristics of the email. You will construct your model using a training data set with information on 57 variables recorded for 919 emails. I've also included an evaluation set consisting of the predictor values (but not the responses) for 309 additional emails. You will use your models to make predictions on these additional emails, which I will compare to their true spam status (that I have recorded, but have withheld) in order to evaluate your model.

You should record your answers in this .Rmd file. However, you are encouraged to use a separate .Rmd file for scratchwork. You will be asked to reflect on your thought process, setbacks, and successes, as well as the code you included in this homework, during our second Code Review week 13, so be sure to keep detailed notes of your investigation.

The assignment is divided into several **Components** to help organize your work. Put all work you want graded between the bars of red stars in the corresponding section.

Grading

Your final score on this assignment will be based 15% on the **performance** of your model predictions on the evaluation data, as measured by accuracy, sensitivity and specificity (as detailed below) and 85% on the quality and depth of your explanations and analysis.

I *do not* expect you to find the absolute best model for this data set. In fact, it is entirely possible to earn top marks on this assignment with a model of mediocre accuracy, provided you submit insightful analysis based on topics we have investigated in our course.

You will be graded as much on your discussion of what *didn't work* and why, as what did.

Goal

At the end of this assignment, you will construct 3 models that you feel best achieve each of the following goals. I've also included benchmarks to help you assess whether you've achieved these goals:

1. The model with the highest overall accuracy (at least 90%)
2. The model with the highest specificity while maintaining reasonable accuracy (specificity at least 95%, accuracy at least 80%)
3. The model with the highest sensitivity while maintaining reasonable accuracy (sensitivity at least 95%, accuracy at least 80%)

Achieving each of these goals will earn at least a B+ on the "performance" component of the assignment.

The Data

The data set `spam_train` can be found in the hw_9 repo and can be loaded by running the following code.

```
spam_train<-read.csv("spam_train.csv")
spam_train$spam <- as.factor(spam_train$spam)
```

Additionally, the `data_description.txt` file in the same repo gives a full description of the variables appearing in the data set.

There is one special column of note:

- `spam` is your response variable (coded as a factor variable, where 1 = spam and 0 = not spam) and should not be included as a predictor.

Components

Data Exploration

In this section, you should perform preliminary data exploration and analysis. This data set is a bit too large to do a full investigation of each variable, so select several variables you think may be useful (at least 6 for visualization purposes, but you may want to use many more for the actual model building). Create visualizations for each variable individually, along with visualizations showing the relationship between the variable and the response. Compute relevant summary statistics for each of your variables.

Model Building

In this section, you should build a series of models (at least 5, but more is probably better) of varying complexity and that use a variety of the tools we have studied thus far. Explain why you choose to implement various features in each model. Here are some suggestions:

- Build at least one model using a large number of variables
- Build at least one model using a small number of variables
- Build at least one highly flexible model
- Build at least one highly rigid model
- Build at least one model that has a parameter that can be tuned using cross-validation
- Consider models using a variety of classification threshold (i.e. the value of c in the rule "Predict $Y = 1$ if $P(Y = 1|X) > c$")

Model Selection

In this section, evaluate model performance using a variety of metrics. Which models seemed to perform better or worse? Why? Here are some suggestions:

- Use cross-validation as well as training + test sets to evaluate performance
- After assessing performance, revisit models and make small changes
- Consider the structure of the predictors. What relationships do these predictors have? What types of models will tend to work best for these relationships?
- Consider modifications that can be made to increase either sensitivity or specificity.

Your model

Goals: Identify the three models you feel will best satisfy the following goals:

1. The model with the highest overall accuracy (at least 90%)
2. The model with the highest specificity while maintaining reasonable accuracy (specificity at least 95%, accuracy at least 80%)
3. The model with the highest sensitivity while maintaining reasonable accuracy (sensitivity at least 95%, accuracy at least 80%)

Load the evaluation data using the following code:

```
spam_eval <-read.csv("spam_eval.csv")
```

Use your 3 models to make 3 sets of predictions on `spam_eval`:

1. Make predictions using the model that best achieves goal 1. Save your predictions as the data frame called `FirstName_LastName_goal1`
2. Make predictions using the model that best achieves goal 2. Save your predictions as the data frame called `FirstName_LastName_goal2`
3. Make predictions using the model that best achieves goal 3. Save your predictions as the data frame called `FirstName_LastName_goal3`

Be sure to replace `FirstName_LastName` with your actual first and last names in the above data frame names.

Once you have made predictions, remove the # symbol from the following lines of code, and replace `FirstName_LastName` with your actual first and last name. **Do not delete the letters .csv at the end of the file name inside the quotations.** Then run the code. This will save your data frames as .csv files in your repo.

```
# write.csv(FirstName_LastName_goal1, "FirstName_LastName_goal1.csv", row_names = F)
# write.csv(FirstName_LastName_goal2, "FirstName_LastName_goal2.csv", row_names = F)
# write.csv(FirstName_LastName_goal3, "FirstName_LastName_goal3.csv", row_names = F)
```

Be sure you commit and push these .csv files to the Github repo in addition to your .Rmd, as they are the predictions I will use to evaluate your model.

Conclusions

Discuss some limitations of your methods and your model. What are some ways you could improve your model if you had more **time**? Identify one variable **not** in the data set you feel could be an important predictor of `spam`. How confident are you in the accuracy of your model?