## Abstract

The use of email has grown exponentially since the introduction of the world wide web in the late 20th century. Today, spam email is ubiquotous on every email platform. Spam detection methods to filter out unwanted emails originated in the late 1990s, and while the algorithms have improved, so have spam avoidance methods. In this paper, we explore rpart; a classification and regression tree package in R. Specifically, we explore the effectiveness of spam classification using rpart and its hyperparameters on a dataset of emails previously classified as spam or valid email. After parameter optimization, our custom rpart model outperforms an rpart model using default settings.

## Introduction

Electronic mail is an integral part of everyday life. The information spread through the use of email is massive in scale, and often includes unwanted marketing and phishing. Email adoption spread rapidly in the 1980s, and so did the ability for companies to market products to email users. In addition, entities with ulterior motives sprang up, attempting to gain information from people using phishing and social engineering. These unwanted emails were given the term “spam.”

Spam filters were introduced not long after the introduction of email. These filters automatically process incoming messages and apply different statistical techniques to identify and remove unwanted emails. Bayesian email filters began to be utilized in 1996, but didn’t become popular until much later. These techniques utilize the probabilities of certain words occurring in regular emails versus spam emails to determine if a message is spam or valid.

Today, many different analytical methods exist for spam email classification. One such spam filtering method is the utilization of decision trees. We explore a decision tree package in R called rpart, which is short for recursive partitioning. Our objective is to investigate and optimize key hyperparameters used in the rpart package in order to classify email messages as spam or valid email. In order to accomplish this, we fit a default decision tree and an optimized decision tree and compare them.

In the subsequent section, we review research literature and introduce the email dataset. In the methods section, we explore the email dataset and explain the methods used for optimizing rpart decision tree parameters. We implement two different decision trees and compare their generalization performance to determine if hyperparameter tuning results in better performance. Our paper concludes with a discussion of the applications of the improvements in anti-spam email filtering.

## Background

In their paper from the Hawaii International Conference on System Sciences, Cukier and Cody define spam as “including all electronic messages that are unsolicited or unwanted, sent to many users (in bulk), without regard to the identity of the individual user, and usually having commercial purposes” (1). Spam also includes messages containing attachments that spread viruses through emails. They state that in 2002, spam numbers peaked at one in three email messages. Also, in 2003, approximately 20 billion spam messages were sent daily.

Clearly, spam has become a major problem for users, businesses, and the internet in general, which led to the introduction of spam filters. These filters have historically relied on keywords within the message to identify spam. Some of the methods used include list-based filters (which classify the sender), content-based filters, challenge/response systems, and collaborative filters where users report spam messages, which are then stored in a database (Jha 2).

Many researchers in academia focus on creating a web spam taxonomy to prevent spam from spreading. The literature explores the different types of spamming and the way it is used to collect information from the users. But as spam detection improves, so does spammers’ techniques to send spam (Takhmiri et. al, 3).

There are many types of classification methods to detect spam. Statistical methods such as support vector machines, naïve Bayes classifiers, and decision trees have commonly been used. The latter is the method used in this project.

A decision tree is a graph used to model outcomes based on a given set of rules. It can be used to classify unlabeled data (Song 4), such as the dataset of emails we use in this case study to identify spam. In our case, we utilize rpart, a recursive partitioning package in R. Recursive partitioning is a statistical method that outputs a decision tree. Regression and classification tasks can be executed using recursive partitioning. The rpart package is one of the most commonly used packages for machine learning in R. It implements the classic non-parametric CART algorithm, using the Gini index as a default splitting criterion. The package also contains a number of optional parameters that can be tuned to optimize model performance. We introduce these parameters in the methods section.

### Data Description

The dataset used for our analysis task is made up of a corpus of emails from SpamAssassin.org (Apache 4). In total, there are 9348 unique emails. It contains 29 predictor variables and one response variable named isSpam. Of the 30 total variables, 17 are boolean factor variables and the remaining 13 variables are numeric variables. Each email has been previously classified as spam or valid. We will use these predictor variables and the isSpam response to create two rpart decision tree models. A high-level listing of variable names is given below.

* Numeric variables*:* perCaps, bodyChartCt, numLines, subExcCt, subQuesCt, numAtt, numRec, hour, perHTML, subBlanks, forwards, avgWordLen, numDlr
* Boolean variables: isSpam, isRe, isYelling, underscore, priority, isinReplyTo, sortedRec, subPunc, multipartText, isPGPsigned, subSpamWords, noHost, numEnd, isOrigMsg, isDear, isWrote

For context, a sample description of five predictor variables is provided in Table 1. We omit the rest of the variable descriptions for brevity and give explanations where appropriate in future sections.

Table 1: Sample Variable Descriptions

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| perCaps | Numeric | Percentage of capitals in the email body |
| isYelling | Boolean | Subject alpha characters are all capital |
| bodyCharCt | Numeric | Number of characters in the body of the email |
| numEnd | Boolean | Email ends in numbers |
| isRe | Boolean | Subject contains reply characters Re: |

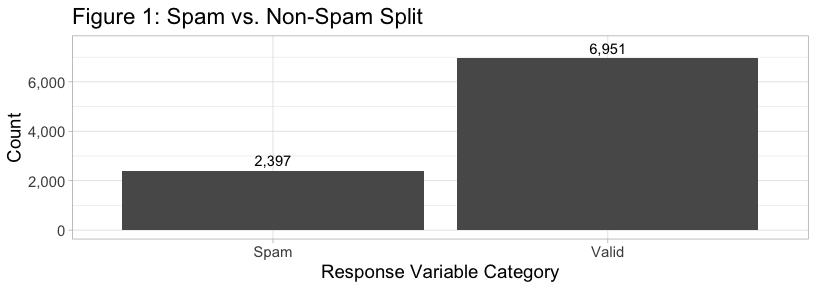
## Methods

Prior to fitting decision trees using rpart, we explore the email dataset in more detail. Specifically, predictor variable relationships are examined using correlation and independence methods. Relationships of the response variable isSpam to predictor variables are also explored. In order to determine variable importance, we use rpart to extract optimal variable splits for a generic decision tree.

### Exploratory Data Analysis

A preliminary evaluation of the dataset found missing observations in 303 unique rows. The availability of values for the following predictor variables was incomplete for seven variables.

Rather than discard over three percent of the dataset, we undertake imputation using random forest regression and classification methods for both numeric and categorical predictor variables.



Our dataset is unbalanced, with 2,371 (26%) spam emails, and 6,674 (74%) valid emails. This imbalance in the dataset could introduce higher false negative rates for our analysis task.

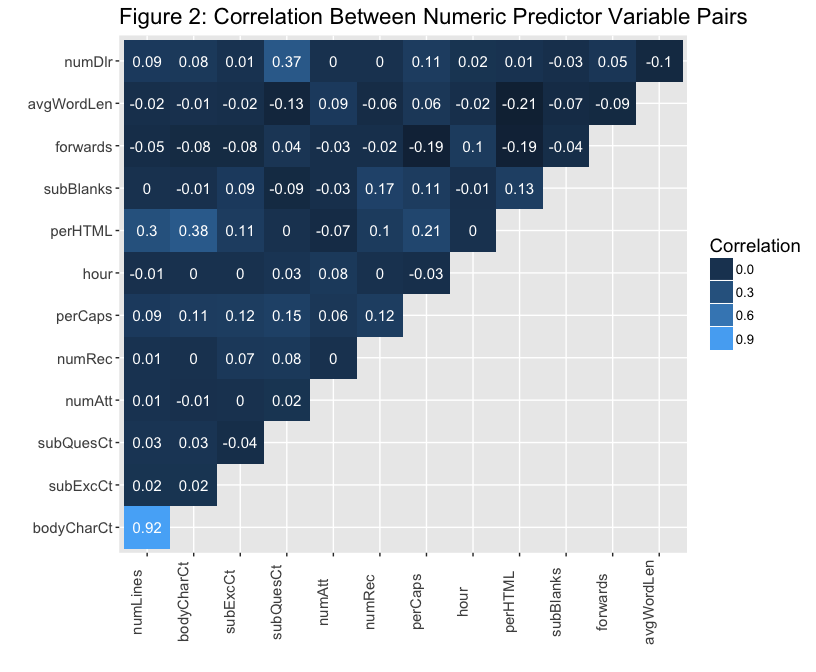
#### Explanatory Variable Relationships

Considering our imputed dataset, the correlation matrix in Figure 2 reveals several significant positive correlations between numerical predictors.

Three positive relationships that stand out are listed below:

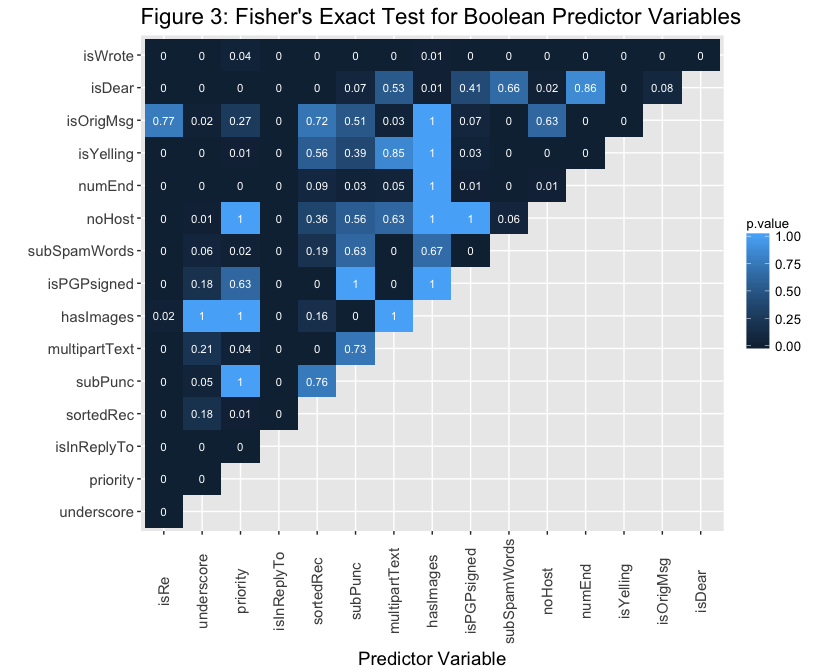
* numLines and bodyCharCt
* perHTML and bodyChartCt
* numDlr and SubQuestCt

The higher correlation between predictor variables could cause these variables to be collectively over-weighted as part of a classification model. Indeed, bodyChartCt and numLines, which represent the number of distinct lines in the body, are nearly similar variables.



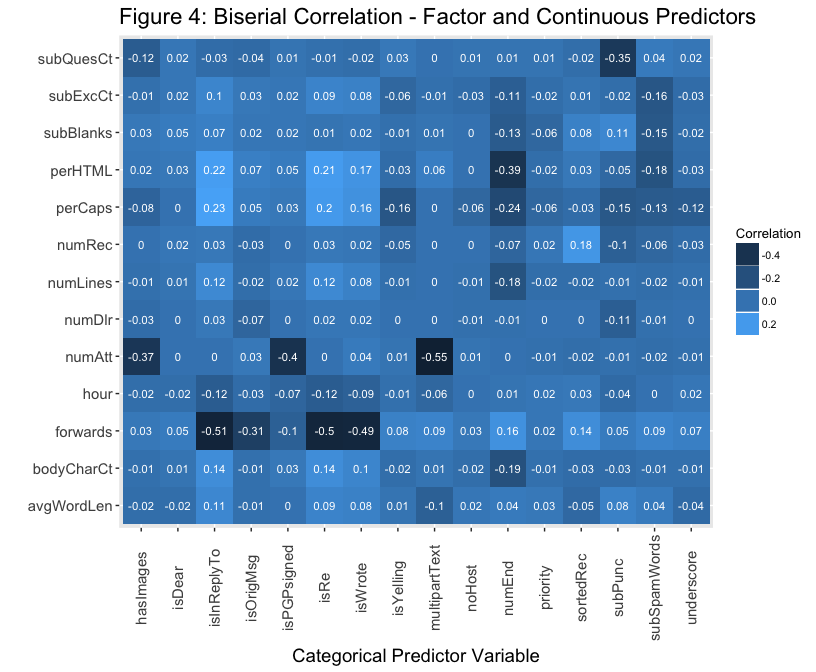
However, given that we are using recursive partioning, collinearity issues are usually rectified naturally. The default splitting algorithm in rpart is greedy and will select the most important variable for classification if similar variables are found.

Given that our dataset contains 16 boolean predictor variables, we display a Fisher’s exact p-value matrix in Figure 3. This figure shows the resulting p-values for dichotomous variables only. Correlation is not an appropriate metric for analyzing nominal or dichotomous relationships. Fisher’s exact test allows us to non-parametrically examine the association between our categorical variables. It also allows us to account for small numbers of observations for certain variable frequency counts.



Lower p-values indicate we reject the null hypothesis of random association. Obviously, significant non-random dependence between factor variables exists. This makes logical sense for variables such as isWrote, which indicates if an email is electronically scribed. Since the majority of emails are electronically scribed, we can assume this variable may not be as important for classifying spam emails. Other interesting relationships such as the independence between priority and noHost indicate that these variables may be useful separately for analysis. The variable noHost indicates a lack of a host name from a sender and priority is set by the sender of the message.

In order to be complete, we also visually inspect the biserial correlation between factors and continuous variables. Given that our factors are all dichotomous nominal variables, biserial correlation is an appropriate measure to use when exploring relationships between nominal and continuous variables. Upon visual inspection, we are able to establish some common sense relationships.



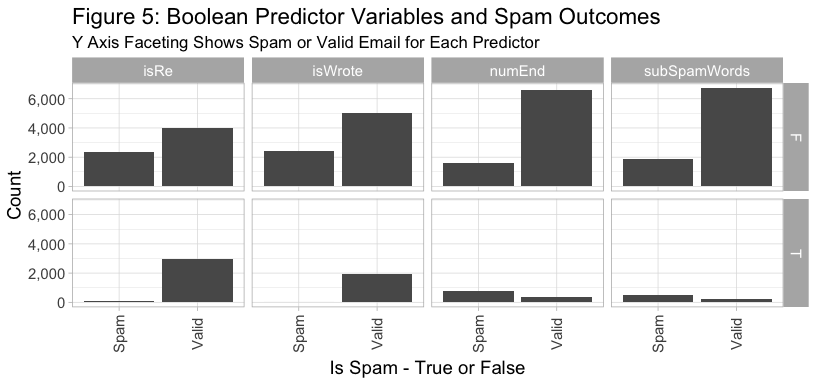
For example, the number of attachments (numAtt) is negatively correlated with the boolean multipartText. Multipart text messages typically do not contain attachments. Additionally, the number of forwards is negatively correlated with isInReplyTo. This makes sense, as replies typically do not contain many forwards. Overall, we see stronger negative relationships when investigating the correlation between factors and continuous variables.

In summary, highly correlated predictor variables could be extraneous to our objective of predicting whether or not an email is spam. To address predictor and response relationships, we visually inspect both continuous and factor variable relationships with the isSpam response variable in the next section. Additionally, we establish variable importances using the rpart package.

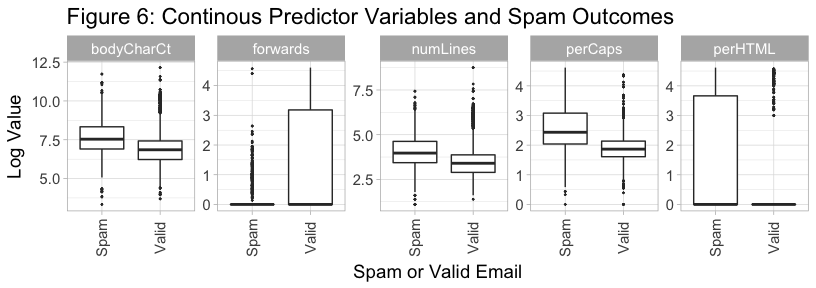
#### Response Variable Relationships

As described earlier, we know that the majority of observations in the overall dataset are classified as spam. However, we can visualize the separation between spam and valid emails for factor and continuous predictors using different plotting techniques.

First, we inspect four factor variables to determine viability for spam prediction: isRe, numEnd, subSpamWords, and isWrote. The variable numEnd indicates whether or not the “from” email prefix ends with a number. For example, “Greer5769@yahoo.com” would be classified as true for numEnd.

The variable subSpamWords is a boolean that is true when a known spam word is contained in the subject. For instance, the word “viagra” would trigger a boolean value of true for subSpamWords. The majority of spam occurs when the four factors above are set to false. On the other hand, we know that mostly valid emails occur when isRe and isWrote are set to true. There are occurences of spam emails in both boolean statuses for factor predictor variables numEnd and subSpamWords. However, the true cases of each predictor variable have more cases of spam than valid email. Given the splits above, we can see how a decision tree could split on different categorical variables in order to classify an email message as spam or valid.

We can also identify separation of spam and valid emails by looking at log values for each numeric predictor in a box plot. We inspect five numeric variables in Figure 6.



The predictor variable forwards, which quantifies how many times an email has been forwarded, shows a more concentrated distribution of values in the third quartile for messages that are valid. The predictor variable perCaps shows a larger interquartile range for spam. We also know from the boxplot that the median perCaps value for spam messages is higher than for valid messages. Also, roughly 75 percent of valid messages have a perCaps value lower than the median perCaps value for spam messages. PerHTML, which represents the percentage of HTML in the email body, also provides for a decent classification variable. The majority of its third quartile occurs specifically with spam predictions.

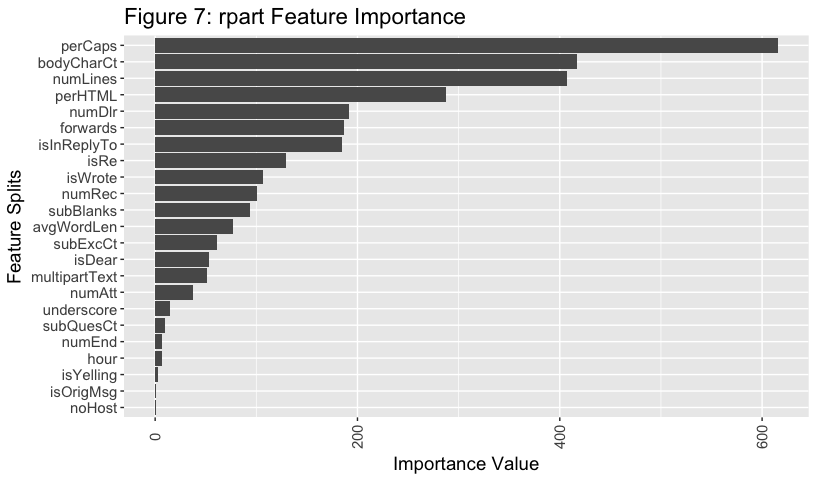
Examination of these predictor variables gives some idea of how we should expect rpart to determine splits. However, we can also expect that some of the lower correlation variables might be involved in the decision of spam versus valid, perhaps providing some finer detailed distinctions between the two classes. We use rpart in the next section to determine which variables are the most important for splitting.

#### Variable Selection and Model Comparison Setup

Instead of fitting complex variable selection algorithms to our email data set, we fit an rpart model on the training data using all 29 features and default model parameters. Default parameters for rpart consist of a minsplit of 20, a complexity paramer (cp) of 0.01 and a maxdepth of 30. The Gini index is used as the splitting criterion by default.

80 percent of the email data is used for training and 20 percent is used for testing. Spam represents roughly 25 percent of the emails in our original dataset. Therefore, we stratify the observations in our training and test data sets to maintain the original spam-to-valid ratio.

We also retain this initial fit on the training data set in order to evaluate the model’s generalization capabilities on the test data set. We call this model our “base” model, and use this terminology going forward. After the rpart model is trained, the package provides us with a listing of the variables it considers the most important for splitting. We can easily identify these variables with the variable importance method contained in rpart. Figure 7 shows the variables our base model considers the most important for classification.

At the top of the variable importance list is perCaps, or the percentage of capital alpha characters in the body. Importance is a weighted sum of improvement in impurity for each variable split. The perCaps variable overshadows all other variables from an importance standpoint. Indeed, we saw good separation of spam and valid email in the previous section for this variable. We are able to classify 77 percent of all spam messages based on a split value of 13 percent for perCaps. In this case, the base model predicts non-spam when a message contains less than 13 percent capitals. Additionally, we see email from addresses ending in numeric (numEnd) and priority provide little value for spam classification. Base model performance and comparisons are given in the results section.

We also investigate the effect rpart’s control parameters have on classifying spam emails by fitting a separate, optimized rpart model. For this model, we analyze four different parameters: complexity penalty (cp), minsplit, maxdepth, and the splitting criteria. Table 2 provides a description for each of these parameters.

Table 2: rpart Hyperparameters

|  |  |  |
| --- | --- | --- |
| Parameter | Description | Default Value |
| Complexity Parameter (cp) | A scaled complexity penalty that ranges from 0 to 1. cp is compared against the error rate relative to a previous split. Splits that do not decrease the lack of fit by cp is not considered. | 0.01 |
| Minsplit | The minimum number of observations that must exist in a node in order for a split to be attempted. | 20 |
| Maxdepth | The maximum depth of any node of the final tree, with the root node counted as depth 0. | 30 |
| Splitting Criterion | Gini or information. Gini utilizes the Gini index to optimize split points, information uses entropy and information gain. | Gini |

Utilizing the full listing of variables in our email dataset can lead to over-fitting; however, decision trees allow us to find the best variables for splitting while pruning extraneous variable splits. Applying the first three parameters in Table 2 typically reduces the size of the final tree. This reduction in size can help with model generalization to the test dataset.

The rpart package contains other control parameters used primarily for exploratory purposes (Atkinson 6). However, one point of clarification regarding the parameter xval is warranted. The parameter xval allows a user to optimize the cost penalty (cp) for a tree with k-fold cross-validation. Given that rpart does not allow for the tuning of multiple parameters simultaneously, we instead rely on the mlr package (Bischl 7) for cross-validation and tuning. The mlr package provides a broad toolset for machine learning tasks in R. It is primarily a wrapper for other machine learning packages, and provides fantastic utilities to automate tedious tasks in a machine learning workflow.

#### Hyperparameter Optimization

Optimization of parameters is achieved by exploring a discrete list of the four parameters of interest. The grid search method is used in conjunction with ten-fold stratified cross-validation.

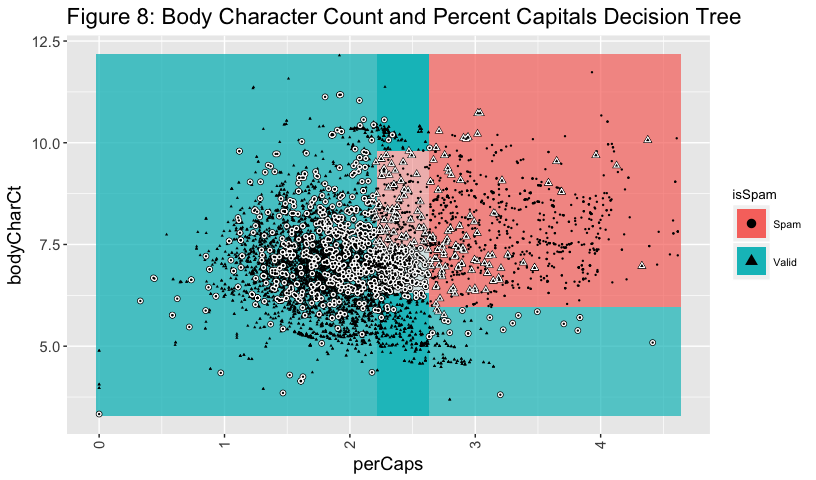
We seek to maximize the true positive classification rate where “spam” is our positive class. We do this by selecting area under the ROC curve (AUC) as our performance metric when tuning hyperparameters. A false positive means an important email may end up in spam or deleted. A false negative means the user may experience unfiltered messages that should be in the spam folder. We consider the former situation a more severe model error. The mean AUCs for all models and associated parameters are compared, and the model with the highest cross-validated AUC is chosen as our optimized model.

The base rpart model with default parameters is subsequently compared to the optimized model given the test data set. We explore model performance and comparisons in the next section. We consider final model comparisons using a holistic set of performance metrics including AUC, mean misclassification error, false positive rate, and false negative rate.

## Results

### Base Model Results

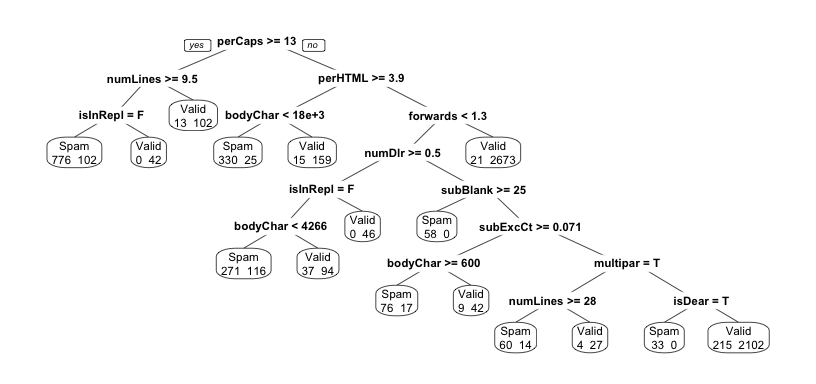
Our base model lists perCaps as the most important variable to split in Figure 7. In second place is BodyCharCt, which represents the number of characters in the body of the email message. Given that these variables have such a higher importance value, we visually inspect an rpart model fit on our training data set using only perCaps and BodyCharCt.



We represent values on a log scale in order to provide a clean visualization of classification regions. The observations outlined in white are misclassifications and the color boundaries represent the outcomes spam and valid. The lighter areas of the chart represent lower probabilities for a given class. Obviously, the ligher regions represent areas of higher misclassification. Additionally, utilizing only these two variables, our base model struggles with false positive rate. This indicates our model needs more complexity. However, these two variables alone provide for good separation as part of a more complex model.

Thus, we leave it up to rpart and its default parameters to create a decision tree containing 14 splits represented in Figure 9. This default rpart decision tree was created in the previous section. We fit this default model on the entire training data set.

Figure 9: Default rpart Decision Tree Representation



Of particular note is the base model’s usage of bodyCharCt. It is used three times in the splitting process for the training data. Additionally, we can see quite a bit of misclassification in the model on the training set, particularly a prevalence of false negatives.

We utilize the test data set of emails to produce a confusion matrix and determine model generalization performance.

Table 3: Confusion Matrix for Default rpart Model

|  |  |  |
| --- | --- | --- |
| **True** | **Spam** | **Valid** |
| **Spam** | 403 | 76 |
| **Valid** | 77 | 1313 |

Table 4: Performance Report for Default rpart Model

|  |  |  |  |
| --- | --- | --- | --- |
| **AUC** | **False Positive Rate** | **False Negative Rate** | **MMCE** |
| 0.95 | 0.06 | 0.16 | 0.08 |

In order to address overall misclassification, we include MMCE, or model misclassification error rate in our metrics set. As seen in the confusion matrix, the base model struggles with false negatives. In total, 76 observations in the test set out of 479 spam records are misclassified for a false negative rate of 15.8 percent. MMCE overall is high at almost eight percent. We attempt to improve on these metrics by optimizing rpart’s hyperparameters.

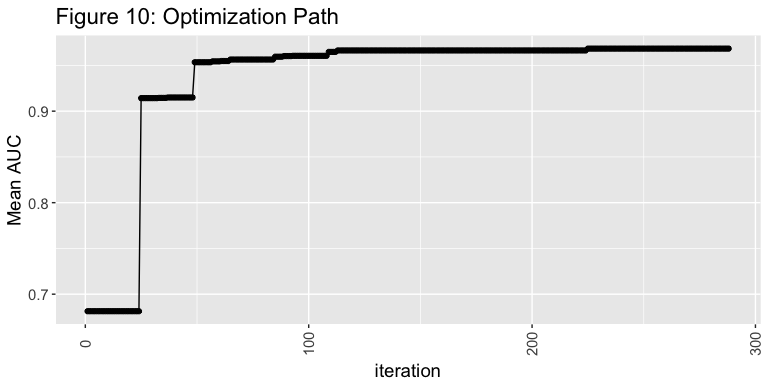
### Optimized Model Results

As stated previously, we explore the complexity parameter, minimum split for each node, maximum depth of the tree, and the splitting criterion via a grid search. The parameter list is given in Table 5 below:

Table 5: Parameter Search Criteria

|  |  |
| --- | --- |
| **Parameter** | **Range** |
| Complexity (cp) | 0.001, 0.01. 0.1, 0.2, 0.5 |
| Minsplit | 1, 5, 10, 15, 20, 30 |
| Maxdepth | 1, 5, 10, 15, 20, 30 |
| Splitting Criterion | Gini, Information |

Each combination of parameters is cross-validated on 10 folds of the training data set using AUC as the performance metric. After the grid search and associated cross-validations are complete, the model with the best AUC performance has a complexity penalty of 0.001, a minimum split of 10, a maximum tree node depth of 15, and uses information as its splitting criterion. We can see the optimization path for our rpart model clearly by viewing the AUC score in sorted order.



Optimization begins to stall just under an AUC of 0.97 for our training data. Our optimized tree is much larger than our base model; 76 total splits are used for our optimized tree. This is because the optimal complexity penalty is set lower than the default (0.01) at 0.001. Such a low cp may introduce risk for overfitting our test set, which we will explore shortly.

Of the 29 models fit during the tuning process with a resulting cross-validated AUC of greater than 0.96, all have complexity parameters of 0.001. Additionally, the maxdepth for a decision tree node is always 10 or higher. 14 out of the 29 models use the information splitting criterion and the majority of models have a minsplit value of 10 or higher. Optimization results favor more splits with medium node depth when compared to the default model. A single unique complexity parameter indicates dominance when it comes to maximizing AUC results for our analysis task. The splitting criterion is nearly a toss up between Gini and information criteria.

We can easily see the complexity parameter’s dominance in Figure 11 below, where the mean AUC across validation splits is maximized as long as the complexity parameter remains small.

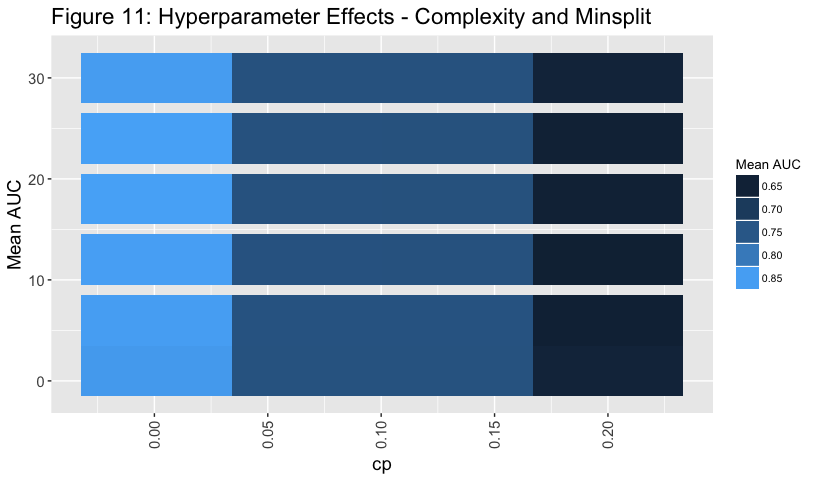
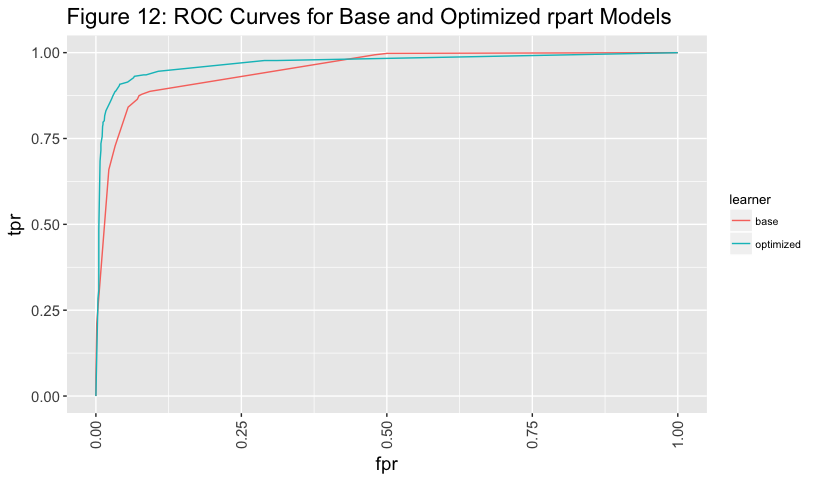
After thoroughly exploring rpart parameters, we turn to generalization performance. Given an AUC of 0.95 for our base model, we fit our optimized model to the same training data set of emails.

Table 6: rpart Model Comparisons

|  |  |  |
| --- | --- | --- |
| Metric | Optimized | Base |
| auc | 0.97 | 0.95 |
| mmce | 0.05 | 0.08 |
| fpr | 0.03 | 0.06 |
| fnr | 0.11 | 0.16 |

Our optimized model outperforms the base model on our key metric: AUC. Given that our base model struggled with false negative rates, we also include false positive and false negative rates for the classification of spam email on the test data set. Our model significantly improves the false negative, false positive, and misclassification error rates as well. However, no analysis using AUC would be complete without viewing an ROC curve for comparison.

We can now visually attest our optimized rpart model outperforms the base rpart model using default parameters. The optimized model contains more area and results in a better true positive to false positive ratio trade off.

## Conclusions and Future Work

Through detailed data exploration, extensive parameter investigation, and tuning, we are able to improve upon rpart’s default settings in an email spam classification setting. We determined that the complexity penalty in rpart dominates when optimizing for AUC given our analysis task. Further, we were able to effectively generalize to a test data set using a relatively complex rpart decision tree.

Additional improvements could be achieved given a more balanced data set. Oversampling of spam emails could reduce the tendency for rpart models to give false negative results. Additionally, empirial investigation of decision thresholds and learning curves could assist in optimizing the classification criteria and amount of training data necessary to fit an optimal rpart model.

Supreme Court Justice Potter Stewart is famously credited with the quotation: “But I know it when I see it.” [1] In context, Justice Stewart was referencing a film that was thought to be outside the standards of decency accepted in the era of the 1960s. In the present day, most internet and mobile phone users could also apply it to the subject of spam, or commercial, unsolicited email: they know it when they see it.

The borders of so-called “spam” email are hard to delineate versus email accepted by users. It would be easy to set extremely broad borders that might include sweeping up any commercial email for disposal before delivery to the recipient. It would be easy as well to be too lenient, filtering only the most egregious examples and leaving the blatant examples of spam to the refuse.

Early efforts at filtering spam were centered on the assumed validity of identity of emailers: recipients were expected to “white list” their favored senders, and the senders validity would be assumed valid if the email was shown to be from the same address. In the last 20 years, many major email handling systems have moved from this type of certainty label to an ensemble of factors that can be labeled with a final “TRUE” or “FALSE” to the email’s spam content indication. Additional aid is given in creation or validation of this level by the input predictor set.

Google, via it’s gmail product, as well as other providers, have had an ample data set over the past several years. All have searched for algorithms that reduce the number of false positives (labeled as spam, when it is actually valid) and false negatives (labeled as valid, when it is actually spam).

A 2012 article in the Journal of Economic Perspectives [2] cites that up to 3% of the 50 billion pieces of spam email sent each today are successful in reaching the recipient. The authors of the paper develop a cost of $20 billion annually to American consumers from spam.

The value of intercepting spam communications, across mediums, is also growing. Most mobile phone numbers have experienced the spoofed-number attack, where a number similar to your own places a call with a robot caller on the opposite end. A number of smart phone applications are currently trying to address this problem by machine learning. Likewise, SMS text messaging spam also provides a risk to users of mobile phones. In both cases, the originator of the call or message is not the number that is listed on the phone, but is a false or spoofed phone number used with the intent of deceiving the recipient. Phishing attacks via SMS skirt the boundaries of security in several ways. First, traditonally there have not been products to address it. Second, the perceived higher level of intimacy by users given to SMS versus email seems to reduce the users’ awareness of threats. Third, the smaller screen may reduce users’ awareness to visiting spoofed websites.

As email filtering continues to move in the direction of machine learning, development of new variables is likely an area for improvement in accuracy and performance. This could move further into the area of natural language processing, and perhaps more specificity to different languages. One can also imagine moving some of the service closer to the user, to further allow for customization. Of course, spam senders will also try to defeat the measures equally as much, and one can imagine training an adversarial model to allow for spam email to pass through filters and land in your inbox. Engineers and developers will need to continue to walk a fine line: avoiding false positives (thereby filtering legitimate emails), while ensuring users’ experience and time is not absorbed by spam emails.

## References

1. Cukier, W.L., and S. Cody. “Genres of Spam: Expectations and Deceptions.” IEEE Conference Publication, 4 Jan. 2006,  
   ieeexplore.ieee.org/document/1579391/.
2. Jha, K. “Comparison of Naïve Bayes Classifier, Decision Tree and an ANN for the Purpose of Spam Detection.” Wisconsin.edu,  
   homepages.cae.wisc.edu/~ece539/fall13/project/Jha\_rpt.pdf.
3. Takhmiri, H., and A. Haroonabadi. “Identifying Valid Email Spam Emails Using Decision Tree.” International Journal of Computer Applications Technology and Research, vol. 5, no. 2, 5 Feb. 2004, ijcat.com/archives/volume5/issue2/ijcatr05021004.pdf.
4. Song, L. “EMAIL SPAM DETECTION USING MACHINE LEARNING.” EMAIL SPAM DETECTION USING MACHINE LEARNING, ese.wustl.edu/ContentFiles/Research/UndergraduateResearch/CompletedProjects/WebPages/sp14/SongSteimle/WebPage/classifiers.html.
5. “SpamAssassin.” Apache, spamassassin.apache.org/.
6. Atkinson, Beth, et al. “Package ‘Rpart’.” Cran.r-Project.org, cran.r-project.org/web/packages/rpart/rpart.pdf.
7. Bischl, Bernd, et al. “Machine Learning in R.” Mlr Tutorial, 7 Mar. 2018, mlr-org.github.io/mlr-tutorial/release/html/.
8. Nolan, D., Temple Lang, D. DATA SCIENCE IN R: a Case Studies Approach to Computational Reasoning and Problem Solving. CRC PRESS, 2017.