

My Comprehensive Evaluation

A Comprehensive Evaluation Report

Presented to
The Statistics Faculty
Amherst College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts
in
Statistics

Ian Fayorsey

November 2018

Acknowledgements

I would like extend my graditude to several people. Professor Nick Horton, and Professor Amy Wagaman, for their guidance, and encouragement over the past 4 years at Amherst. Since declaring my major sophomore year, Professor Horton has been my advisor and confidant, helping me through any and all adversities. During challenging times, Professor Wagaman helped me remain confident and hopeful. My academic development from probability to data analysis is a result of your unwavering support. To the entire statistics department, thank you for equipping me with the skills and thought process to tackle future problems in a formulaic and interpretable manner. Finally, to my mother, father, and sister - for never allowing me to accept failure and helping me become the man I am today.

Table of Contents

Introduction	1
0.0.1 Previous Work	2
Chapter 1: Data	3
1.0.1 Data Preprocessing	3
1.1 Exploratory Data Analysis:	4
Chapter 2: Theoretical Background	9
2.1 Support Vector Machines:	9
2.2 Decision Trees	11
2.3 Bagging	12
2.4 Random Forests	13
2.5 Linear Discriminant Analysis	13
Chapter 3: Results	25
3.0.1 Run Time	25
3.0.2 Classification Accuracy	26
Conclusion	29
Appendix A: Appendix	31
References	41

List of Tables

List of Figures

1.1	Pie Chart	5
1.2	Word Count Distribution	6
1.3	TD-IDF per Category	6
1.4	Framework of Text Classification	7
3.1	Bar Plot of Missclasification Rate	26

Abstract

A collection of unstructured data is analyzed using six machine learning algorithms, that classify responses to pre-defined categories. Performance of these models is evaluated in terms of misclassification rate and learning speed. It was discovered that simple machine learning algorithms are more computational efficient than complex models. Support Vector Machines were the most efficient method employed. We conclude by examining the factors that make text classification challenging.

Introduction

By 2022, 93% of all data in the digital universe will be unstructured (Big Data, 1). As the wealth of information stored in comments, tweets, and reviews increases, so does the growing interest in extracting insight from this unstructured data. Organizations such as Amazon, Apple, and Facebook have begun recruiting researchers to investigate how data can be used to better reach their customers. Over the years, this research has led to break throughs in text categorization – the assignment of natural language texts to predefined categories based on content – which has applications in areas such as search engines and customer relationship management. Some of the biggest challenges when working with unstructured data pertain to the volume of data. By nature, a large percentage of the data companies collect is unverified, and remains in its uncleaned, user generated state. When drawing insights from this data, it is vital that the data is transformed into an actionable form. In the case of analyzing text, there are 171,476 words in the English language, making natural language processing a difficult task when dealing with its large volume. In this paper, I look to find the most computationally efficient methods of analyzing natural language.

Six supervised methods were used in this analysis: Support Vector Machines (SVM), Multinomial Logistic Regression, Decision Trees, Bagging, Random Forest, and Linear Discriminant Analysis (LDA). Misclassification rates, precision-recall, and runtimes were used to evaluate each algorithm's computational efficiency. Misclassification rate is defined as 1 minus classification accuracy (the proportion of correctly predicted categories). The report is organized as follows:

- Chapter 1 provides background information on the chosen dataset, how it was mined, how it was transformed, and an exploratory analysis of the features within the data
- Chapter 2 provides the theoretical background of each learning algorithm used in the experiment
- Chapter 3 shows and interprets the results, and concludes

0.0.1 Previous Work

My exposition draws from several noteworthy analyses of textual data. This list includes the role of SVMs in learning text classifiers (Joachim, 1994), the use of semantic orientation in classifying documents (Turney, 2002; Pang et al., 2002), and the benefits of term-frequency transformations (Leopold and Kinderman, 2002).

Chapter 1

Data

The data used in this analysis is a series happy moments. In collaboration with the University of Tokyo, Massachusetts Institute of Technology researchers interview thousands of people and asked them to list ten happy moments that occurred within the last 24 hours. Their responses were recorded and compiled into *HappyDB*, a corpus of 100,535 happy moments. The data set is host publicly on github, as a zip file (<https://rit-public.github.io/HappyDB/>).

It is important to note that this curated data set contains cleaned textual data. In it includes 9 variables, many of which identify the author and qualities of the texts. The variables of interest in this analysis are `cleaned_hm`, the list of happy moments, and its associated predicted category. For my analysis, I sampled 15,000 of these responses and trained on 80% of the data.

Response	Predicted Category
I was happy when my son got 90% marks in his examination	Acheivement
I went to the gym this morning and did yoga	Excercise
My dad and I went fishing	Bonding

1.0.1 Data Preprocessing

In order to analyze how words influence a predicted category, the data set had to be transformed. First, each character vector response is converted into a corpus. A corpus is simply a collection of natural language constructed with a specific purpose. The corpus consists of 15,000 responses samples from the larger dataset.

```
as.character(corpus[[1]])
```

```
[1] "i went to the farm store and found fruit trees on sale for 70% off. "
```

Entries in the corpus are then cleaned of symbols, punctuations, and stop words. Once accomplished the corpus is reduced a collection of key words.

```
as.character(a[[1]])
```

```
[1] "    farm store    fruit trees    sale    "
```

The next step is to create a document term matrix. A document term matrix is a corpus transformation that represents each word as a feature, and each response as a row. The matrix is populated with values 0 or 1, depending on the absence or presence of that word in the document. Traditionally, document term matrices are high dimensional due to thousands of words that can be used as a feature. Often times many of these cells are empty to represent the absence of a word from that document. High sparsity (the proportion of entries which are zero) is another known condition of document terms matrices.

	Terms			
Docs	farm	fruit	sale	store
1	1	1	1	1
2	0	0	0	0
3	0	0	0	0

1.1 Exploratory Data Analysis:

In this section, I present some preliminary results from summaries and exploratory analysis of the cleaned data obtained above. First we look at the response variable **predicted category**. Next, I explore statistics regarding the explanatory variable **words**.

Predicted Category

Affection and achievement were the most talked about categories (33.9% and 34.1% respectively). This follows conventional logic given that love and sense of accomplishment are vital traits to a good quality of life. Nature and exercises accounted for just 1.2% and 1.8% of labeled responses in the entire data set.

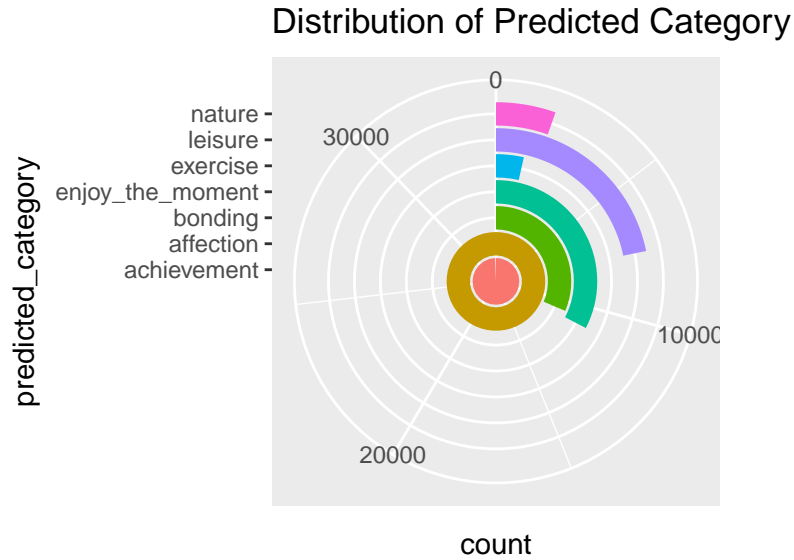


Figure 1.1: Pie Chart

In a predictive setting, having highly disproportionate classes can lead to errors when making predictions based on that training data. When classes are unequally represented, models may simply decide to always predict a certain class in order to achieve a high accuracy. This accuracy paradox does not reflect effective predictions for the model, but rather the state of the underlying distribution. While this is not the case for our data set, consideration must be taken for the classifications of minority labels.

Term Count

In Figure 1.2, the majority of entries are between 5 and 14 words. Over 4,000 reviews contain over 50 words. Many of these words add little to no value to an algorithmic interpretation of the sentence. These are the commonly used words such as “a”, “I”, “was” that were removed from the corpus.

Terms used in each category

Two metrics frequently used to quantify the importance of a word are its term frequency (tf), and inverse document frequency (idf). Term frequency measures how frequently a word appears in a document, while idf weights frequently used words less than words rarely used (Leopold, 2002). When combined, the tf-idf is the frequency of a term adjusted by how rarely it is used.

Figure 1.2 shows a breakdown of the highest tf-idf word per category yields further

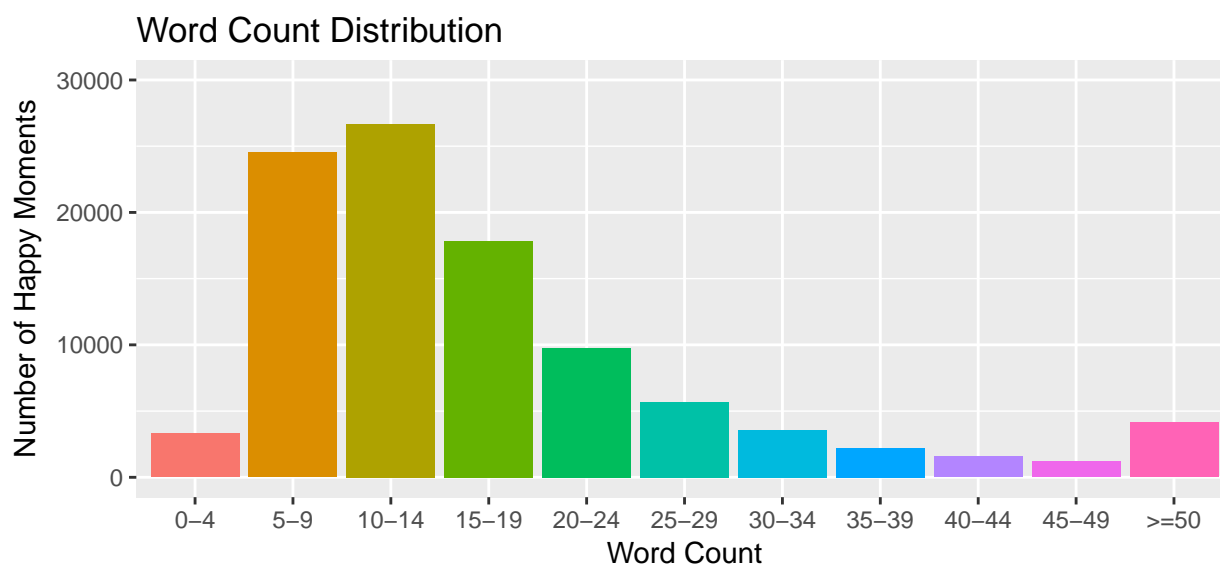


Figure 1.2: Word Count Distribution

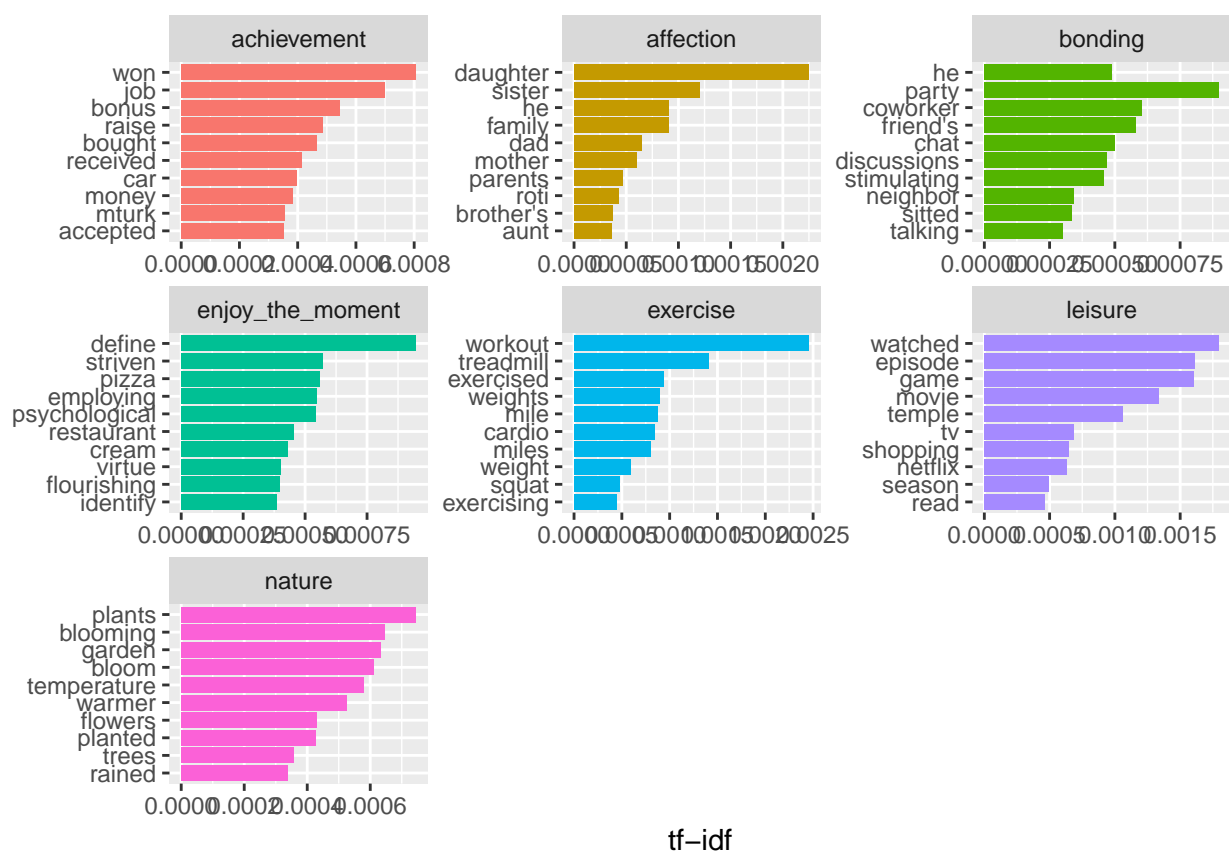


Figure 1.3: TD-IDF per Category

insight into words associated with each label. For phrases related to achievement, words like “won”, “received”, and “accepted” have the highest adjusted usage. In responses categorized as exercise, we see terms related to working out and equipment used. These key terms provide natural separators by which the algorithms can use to classify future responses. In the next section, I discuss some of the methods used to predict the categories of happy moments.

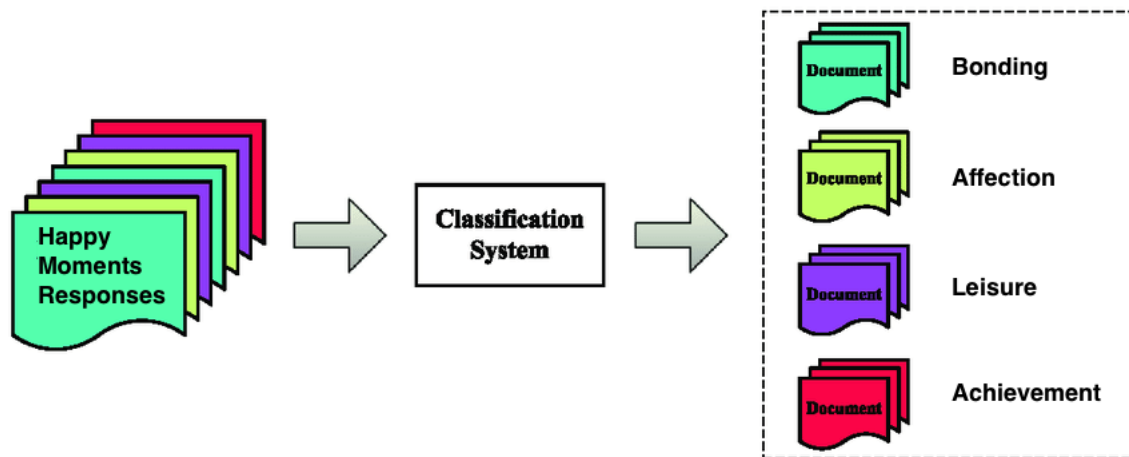


Figure 1.4: Framework of Text Classification

```
minitrain <- dplyr::sample_frac(iris, .7)
minitest <- setdiff(iris,minitrain)
```


Chapter 2

Theoretical Background

Broadly speaking, there are two types of machine learning techniques: supervised, and unsupervised methods. In a supervised setting, inputs are mapped to a predefined label, and models are trained to predict labels for a new set of inputs. In the case of text classification, each response and associated category are learned by a classifier which then predicts labels for new responses. In an unsupervised setting, there is no corresponding labels to the inputs. These methods cluster response based on associations between responses. Given that this study aims to compare the performance of classifiers when labeling texts, we remain in a supervised setting.

The objective is to minimize the classification error rate. The classification error rate is defined as the proportion of training observations in the region that do not belong to the most common class.

2.1 Support Vector Machines:

The Support Vector Machines (SVM) approach to classification is a generalization of the maximal margin classifier to non-linear decision boundaries. It is first defined by a hyperplane, which is a flat subspace existing in a $p - 1$ dimensional setting. The mathematical definition of a hyperplanes is a line with:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$$

The hyperplane essentially divides a p -dimensional space into two halves depending on whether or not the linear equation is greater than zero or less than zero for a given $X = (X_1, X_2, \dots, X_p)^T$. In the scenario where X is an $n * p$ matrix where observations fall into two classes y_1, y_2, \dots, y_n between $-1, 1$. If a separating hyperlane exists, it can

be used to construe an intuitive classifier based on the $G(x) = \text{sign}[x_i^T \beta + \beta_0]$. The optimization problem

$$\max M$$

subject to

$$y_i(x_i^T \beta + \beta_0) \geq M, i = 1, \dots, N$$

Where M is the width of the margin.

In a high dimensional setting however, this is unlikely to be the case. A soft margin classifier aims to address this by allowing some observations to follow on the wrong side of the margin or even hyperplane. These observations correspond to training points that have been misclassified by the support vector classifier. This constraint can now be modified to be:

$$y_i(x_i^T \beta + \beta_0) \geq M - \xi$$

.

Where ξ is a slack variable representing observations on the wrong side of the margin/hyperplane. In this setting, points on the wrong side of the margin are penalized proportionally to the distance from the boundary.

Tuning parameter, C can be introduced restraining the number of points that violate the hyperplane. As C increases we allow for more observations to appear on the wrong side of the hyperplane so the margin widens. Conversely, as C decreases less violations result in a smaller margin. Increasing the cost parameter additionally makes models more expensive and increases the risk of losing model generability.

In situations where there are more than $k > 2$ classes, k SVMs are fit and compared to the remaining $k-1$ classes. Observations are assigned to a class based on which $\beta_0 k + \beta_{1k} X_{1k} + \dots + \beta_{pk} X_{pk}$ is the greatest (Hastie, 337-356).

SVMs hold a unique property in that they can classify observations independent of the dimensionality of the feature space (Joachim). By separating the data on the largest margin rather than the number of features, our model can be easily generalized, irrespective of the size of the feature space.

The issue of non-linearity is typically addressed by enlarging the feature space by using higher order polynomial functions of the predictors.

2.2 Decision Trees

Decision trees partition a feature space into a number of simpler regions. Decision trees are composed of two main parts: a node and leaf. Each node in a decision tree represents a variable by which observations can be group based on conditions associated with that variable. In this setting, every node is a word in the document term matrix. In a classification setting, the tree is split based on which variable minimizes the classification error rate. If we let $\hat{p}_{mk} = 1/N_m * \sum I(y_i = k)$ be the proportion of class k observations in node m , observations are classified to the class where $k(m) = \underset{k}{\operatorname{argmax}} \hat{p}_{mk}$ (Hastie, 312). This can also be described as the majority class for node m . This greedy algorithm chooses which variable to split on based on the largest drop in misclassification rate.

Figure **x** is a simple decision tree that fits chooses petal length and width as nodes. In this instance, all observations in the test set were correctly guessed.

Do to their high variance, resulting trees can vary greatly if the training data is modified, and can lead to drastically different test labels. Traditionally, classification and regression decision trees are associated with high variance and low bias.

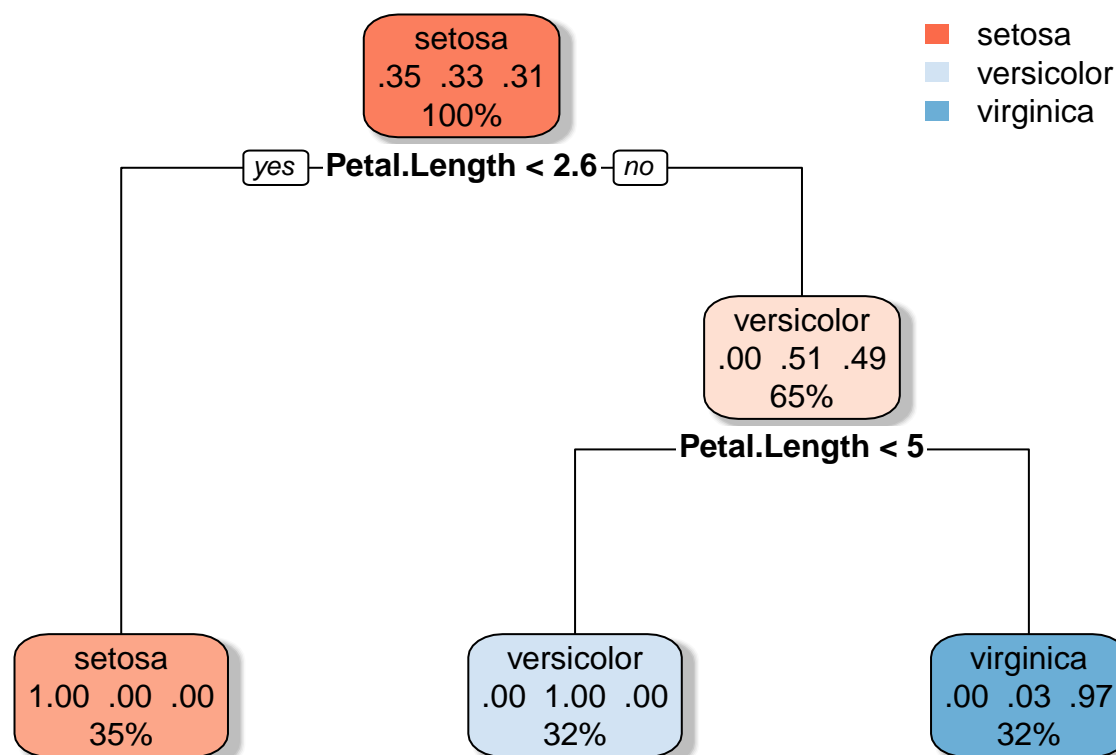
```
tree <- rpart::rpart(Species ~., data=minitrain, method = "class")

tree_pred <- predict(tree, minitest, type="class")

knitr::kable(table(tree_pred, minitest$Species))
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	6
virginica	0	2	43

```
rpart.plot::rpart.plot(tree, box.palette="RdBu", shadow.col="gray")
```



2.3 Bagging

Bootstrap aggregation (bagging) is an ensemble method that combines the predictions of several smaller algorithms to make an improved estimate. Given a set of independent observations $X_1 \dots X_n$ each with variance σ^2 , the variance of the mean \bar{X} is $\frac{\sigma^2}{n}$. This illustrates how averaging over a set of observations can reduce variance. In this approach, several samples are taken from the training data with replacement and a classification tree is trained on each. Given a new data set, the bagging method aggregates the average prediction across the models for each observation and assigns it to the class with the majority vote (Hastie, 317).

Because the prediction is an average of several trees, we are less concerned with one tree overfitting. This allows for trees to remain unpruned, with high variance. In this setting the only adjustable parameter is the number of samples drawn, in essence, how many decision trees are averaged. Traditionally, this number is chosen through cross-validation (increased until accuracy stops showing improvement). It must be kept in consideration however that more samples will require more time to train the models, and may be computationally infeasible.

Due to the algorithm's greedy nature, a series of decision trees may have highly correlated predictions as a result of the method dividing the data based on the largest

drop in missclassification rate. Imagine a scenario where there is only one strongly correlated predictor. Even with a high number of samples, we can expect most trees generated to split the first node on this variable. Situations where there is structural similarity between bootstrap samples can lead to predictions that are highly correlated.

Note: The Bagging model did not provide reproducible results, therefore it is omitted from the remainder of this experiment. Nonetheless, it is an important algorithm that establishes the framework for the next method discussed in this chapter.

2.4 Random Forests

Random Forests algorithms improve on the bootstrap aggregation by decorrelating trees. The main difference is that in a random forest setting, only a random sample of m predictors from p variables can be chosen as a splitting feature. This alleviates the issue described earlier by not considering the one strong predictor in approximately $(p - m)/p$ of the splits. Situations where $m = p$ is simply a bootstrap aggregation (Hastie, 319). Smaller values of m are traditionally helpful when there is a large number of correlated features. Continuing with the iris data set, the random forests model improved on the decision tree estimate ()

2.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is most commonly used as a dimension reduction technique. In the case of $p > 1$ classes, each class density is assumed to follow a multivariate gaussian distribution $N(\mu_k, \Sigma)$, where μ_k is a class mean vector, and Σ is common covariance matrix.

$$\begin{aligned} \log(Pr(G = k|X = x)/Pr(G = l|X = x)) = \\ \log(f_k(x)/f_l(x)) + \log(\pi_k/\pi_l) \\ \log(\pi_k/\pi_l) - 1/2(\mu_k - \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) \end{aligned}$$

When simplified further, the above becomes:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - 1/2 \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

This equation assigns an observation to the class for which its $\delta_k(x)$ is the largest (Hastie, 143). Since the parameters of the gaussian are unknown, following estimates

are used:

$$\hat{\pi}_k = N_k/N \text{ where } N_k \text{ is the number of class-}k \text{ observations}$$

$$\hat{\mu}_k = \sum_{g=k} x_i / N_k$$

$$\hat{\Sigma}_k = \sum_{k=1}^K \frac{1}{N_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$

\$class

[1]	setosa	setosa	setosa	setosa	setosa	setosa
[7]	setosa	setosa	setosa	setosa	setosa	setosa
[13]	setosa	setosa	setosa	setosa	setosa	setosa
[19]	setosa	setosa	setosa	setosa	setosa	setosa
[25]	setosa	setosa	setosa	setosa	setosa	setosa
[31]	setosa	setosa	setosa	setosa	setosa	setosa
[37]	setosa	setosa	setosa	setosa	setosa	setosa
[43]	setosa	setosa	setosa	setosa	setosa	setosa
[49]	setosa	setosa	versicolor	versicolor	versicolor	versicolor
[55]	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
[61]	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
[67]	versicolor	versicolor	versicolor	versicolor	virginica	versicolor
[73]	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
[79]	versicolor	versicolor	versicolor	versicolor	versicolor	virginica
[85]	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
[91]	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
[97]	versicolor	versicolor	versicolor	versicolor	virginica	virginica
[103]	virginica	virginica	virginica	virginica	virginica	virginica
[109]	virginica	virginica	virginica	virginica	virginica	virginica
[115]	virginica	virginica	virginica	virginica	virginica	virginica
[121]	virginica	virginica	virginica	virginica	virginica	virginica
[127]	virginica	virginica	virginica	virginica	virginica	virginica
[133]	virginica	versicolor	virginica	virginica	virginica	virginica
[139]	virginica	virginica	virginica	virginica	virginica	virginica
[145]	virginica	virginica	virginica	virginica	virginica	

Levels: setosa versicolor virginica

\$posterior

	setosa	versicolor	virginica
1	1.000000e+00	2.755892e-24	1.999804e-50
2	1.000000e+00	1.258307e-20	7.343643e-46

3	1.000000e+00	4.111867e-22	9.886500e-48
4	1.000000e+00	7.524052e-19	3.298075e-43
5	1.000000e+00	1.861465e-24	1.523840e-50
6	1.000000e+00	2.894761e-22	1.440340e-46
7	1.000000e+00	1.134807e-20	3.356795e-45
8	1.000000e+00	3.611952e-22	1.546794e-47
9	1.000000e+00	4.749109e-18	2.459414e-42
10	1.000000e+00	4.319250e-21	1.302589e-46
11	1.000000e+00	1.733932e-25	7.254449e-52
12	1.000000e+00	3.197528e-20	9.116529e-45
13	1.000000e+00	2.904170e-21	5.416345e-47
14	1.000000e+00	1.080367e-22	4.660659e-49
15	1.000000e+00	3.058603e-32	4.240894e-61
16	1.000000e+00	1.291673e-28	5.704044e-55
17	1.000000e+00	7.618696e-27	4.162679e-53
18	1.000000e+00	3.006744e-23	1.002316e-48
19	1.000000e+00	6.195071e-24	2.759762e-49
20	1.000000e+00	8.102117e-24	3.786149e-49
21	1.000000e+00	1.751173e-21	1.540744e-46
22	1.000000e+00	3.295413e-22	9.206085e-47
23	1.000000e+00	1.969707e-27	8.224236e-55
24	1.000000e+00	1.353687e-16	4.754209e-39
25	1.000000e+00	8.701887e-17	7.313891e-40
26	1.000000e+00	9.740538e-19	3.695264e-43
27	1.000000e+00	6.002658e-19	1.675869e-42
28	1.000000e+00	1.528008e-23	2.333181e-49
29	1.000000e+00	4.080088e-24	2.624433e-50
30	1.000000e+00	1.119021e-18	7.931613e-43
31	1.000000e+00	1.656706e-18	1.040901e-42
32	1.000000e+00	1.069379e-21	2.080746e-46
33	1.000000e+00	5.217100e-28	3.570846e-55
34	1.000000e+00	6.849271e-30	1.693243e-57
35	1.000000e+00	4.712405e-20	6.528667e-45
36	1.000000e+00	1.844577e-24	4.537677e-51
37	1.000000e+00	4.909648e-27	2.482941e-54
38	1.000000e+00	4.296258e-25	1.123917e-51

39	1.000000e+00	9.124355e-20	1.175435e-44
40	1.000000e+00	1.434406e-22	4.184278e-48
41	1.000000e+00	5.422920e-24	8.591001e-50
42	1.000000e+00	3.956451e-15	1.007915e-38
43	1.000000e+00	6.565203e-21	4.994345e-46
44	1.000000e+00	1.916616e-17	8.677897e-40
45	1.000000e+00	3.358651e-18	6.566091e-41
46	1.000000e+00	3.456930e-19	1.360632e-43
47	1.000000e+00	1.036805e-23	3.258018e-49
48	1.000000e+00	1.445579e-20	1.576259e-45
49	1.000000e+00	4.366182e-25	2.681739e-51
50	1.000000e+00	9.644638e-23	1.739881e-48
51	1.693650e-21	9.999860e-01	1.401777e-05
52	1.186899e-22	9.999325e-01	6.754799e-05
53	8.473702e-26	9.988274e-01	1.172637e-03
54	1.323899e-23	9.999961e-01	3.856578e-06
55	1.107886e-25	9.995925e-01	4.074689e-04
56	1.139350e-25	9.998651e-01	1.348793e-04
57	8.235722e-26	9.966702e-01	3.329779e-03
58	2.599857e-14	1.000000e+00	1.139845e-10
59	1.238463e-22	9.999899e-01	1.011058e-05
60	2.049517e-22	9.999937e-01	6.327598e-06
61	1.738782e-18	1.000000e+00	2.124788e-09
62	2.295770e-22	9.999735e-01	2.645988e-05
63	4.669078e-19	1.000000e+00	7.591480e-09
64	8.023183e-27	9.990226e-01	9.774396e-04
65	3.400325e-15	1.000000e+00	5.940234e-09
66	7.743582e-20	9.999980e-01	1.957826e-06
67	5.270612e-27	9.975381e-01	2.461930e-03
68	3.797901e-18	1.000000e+00	1.354366e-08
69	3.602917e-29	9.979766e-01	2.023377e-03
70	7.699844e-19	1.000000e+00	2.379521e-08
71	1.432055e-31	4.316851e-01	5.683149e-01
72	2.430300e-18	9.999999e-01	1.032369e-07
73	1.172729e-31	9.460091e-01	5.399087e-02
74	2.564111e-25	9.999397e-01	6.032614e-05

75	5.315531e-20	9.999993e-01	7.391705e-07
76	8.248869e-21	9.999963e-01	3.740233e-06
77	9.902233e-26	9.997324e-01	2.675626e-04
78	1.777914e-30	8.232772e-01	1.767228e-01
79	5.694207e-26	9.993090e-01	6.909709e-04
80	2.996368e-12	1.000000e+00	2.977292e-11
81	1.145166e-18	1.000000e+00	1.471550e-08
82	1.744358e-16	1.000000e+00	1.036936e-09
83	6.219294e-18	1.000000e+00	2.995173e-08
84	8.941455e-36	1.762224e-01	8.237776e-01
85	8.288735e-28	9.947091e-01	5.290867e-03
86	3.757635e-24	9.991495e-01	8.504687e-04
87	2.607319e-24	9.997349e-01	2.650713e-04
88	5.632162e-25	9.999837e-01	1.627864e-05
89	2.389636e-20	9.999987e-01	1.284162e-06
90	1.839965e-22	9.999977e-01	2.277387e-06
91	1.969585e-25	9.999653e-01	3.469098e-05
92	4.179032e-25	9.997567e-01	2.433246e-04
93	1.194898e-19	9.999999e-01	1.204051e-07
94	1.756071e-14	1.000000e+00	1.010384e-10
95	3.303425e-23	9.999913e-01	8.741895e-06
96	4.702221e-20	9.999994e-01	5.882133e-07
97	1.156087e-21	9.999965e-01	3.516410e-06
98	8.383142e-21	9.999984e-01	1.593048e-06
99	1.532959e-10	1.000000e+00	6.333465e-12
100	4.329596e-21	9.999985e-01	1.481301e-06
101	6.935569e-60	1.407894e-10	1.000000e+00
102	6.301995e-42	1.022759e-03	9.989772e-01
103	5.765531e-49	1.896401e-06	9.999981e-01
104	3.441684e-44	1.930166e-04	9.998070e-01
105	1.944123e-52	1.273897e-07	9.999999e-01
106	1.433802e-57	4.816846e-09	1.000000e+00
107	2.490032e-35	2.593074e-01	7.406926e-01
108	5.909660e-50	3.343645e-06	9.999967e-01
109	6.238822e-48	3.276585e-05	9.999672e-01
110	2.366994e-54	3.180971e-09	1.000000e+00

111	3.152198e-36	1.207453e-02	9.879255e-01
112	8.645345e-42	1.072817e-03	9.989272e-01
113	3.948866e-44	5.458086e-05	9.999454e-01
114	6.237721e-44	2.768511e-04	9.997231e-01
115	9.676001e-50	6.511572e-07	9.999993e-01
116	3.685324e-45	8.998519e-06	9.999910e-01
117	9.826210e-41	1.669784e-03	9.983302e-01
118	7.523000e-55	4.097757e-09	1.000000e+00
119	4.747967e-68	3.963577e-12	1.000000e+00
120	4.814710e-36	4.485810e-01	5.514190e-01
121	7.352708e-49	6.737926e-07	9.999993e-01
122	8.300989e-41	1.282496e-03	9.987175e-01
123	2.617131e-58	6.209863e-09	1.000000e+00
124	3.108751e-34	2.343325e-01	7.656675e-01
125	6.557139e-46	8.585926e-06	9.999914e-01
126	1.464154e-43	1.479395e-04	9.998521e-01
127	1.250282e-32	4.559468e-01	5.440532e-01
128	2.583794e-33	2.383487e-01	7.616513e-01
129	2.083311e-49	2.246404e-06	9.999978e-01
130	2.857168e-38	1.729229e-02	9.827077e-01
131	1.671249e-48	7.835641e-06	9.999922e-01
132	2.071888e-45	5.494324e-06	9.999945e-01
133	4.156594e-51	4.889965e-07	9.999995e-01
134	2.642103e-32	8.018299e-01	1.981701e-01
135	1.367501e-40	1.778216e-02	9.822178e-01
136	3.148703e-52	9.426251e-08	9.999999e-01
137	5.835140e-51	7.664711e-08	9.999999e-01
138	1.289782e-40	1.480414e-03	9.985196e-01
139	2.386425e-32	3.970447e-01	6.029553e-01
140	3.053523e-41	3.220249e-04	9.996780e-01
141	9.543754e-51	1.615533e-07	9.999998e-01
142	9.750574e-40	4.497674e-04	9.995502e-01
144	1.069275e-52	4.809548e-08	1.000000e+00
145	1.039080e-52	1.927787e-08	1.000000e+00
146	3.411505e-43	5.193417e-05	9.999481e-01
147	7.880243e-39	1.257498e-02	9.874250e-01

```

148 3.136049e-39 2.330916e-03 9.976691e-01
149 1.471640e-46 2.288840e-06 9.999977e-01
150 1.314539e-37 1.498766e-02 9.850123e-01

```

```

$x
      LD1      LD2
1    8.36603369 -0.3407549756
2    7.62664606  0.7701182142
3    7.92910337  0.3419595940
4    7.21122284  0.7410885990
5    8.38822769 -0.5156602667
6    7.78467706 -1.4682568714
7    7.54302483 -0.1765558251
8    7.90622248 -0.0199740083
9    7.06580836  1.1476064200
10   7.74271585  0.8530861971
11   8.60122212 -0.7810366157
12   7.46860526  0.1259016678
13   7.79904327  1.0171710887
14   8.12438146  0.9509460708
15  10.07042971 -1.7771433365
16   9.12997728 -2.7695899559
17   8.81695433 -1.7816490228
18   8.10386253 -0.5540440310
19   8.20420213 -1.1475901630
20   8.18169705 -1.1017033240
21   7.74917965  0.0016667909
22   7.80755795 -1.1063232691
23   9.06140915 -0.6939971416
24   6.58137638 -0.3282398077
25   6.69439731  0.3609457813
26   7.20028138  0.8930504707
27   7.12381085 -0.3682040813
28   8.19773833 -0.2961707568
29   8.34383970 -0.1658496844
30   7.15489542  0.5770037075

```

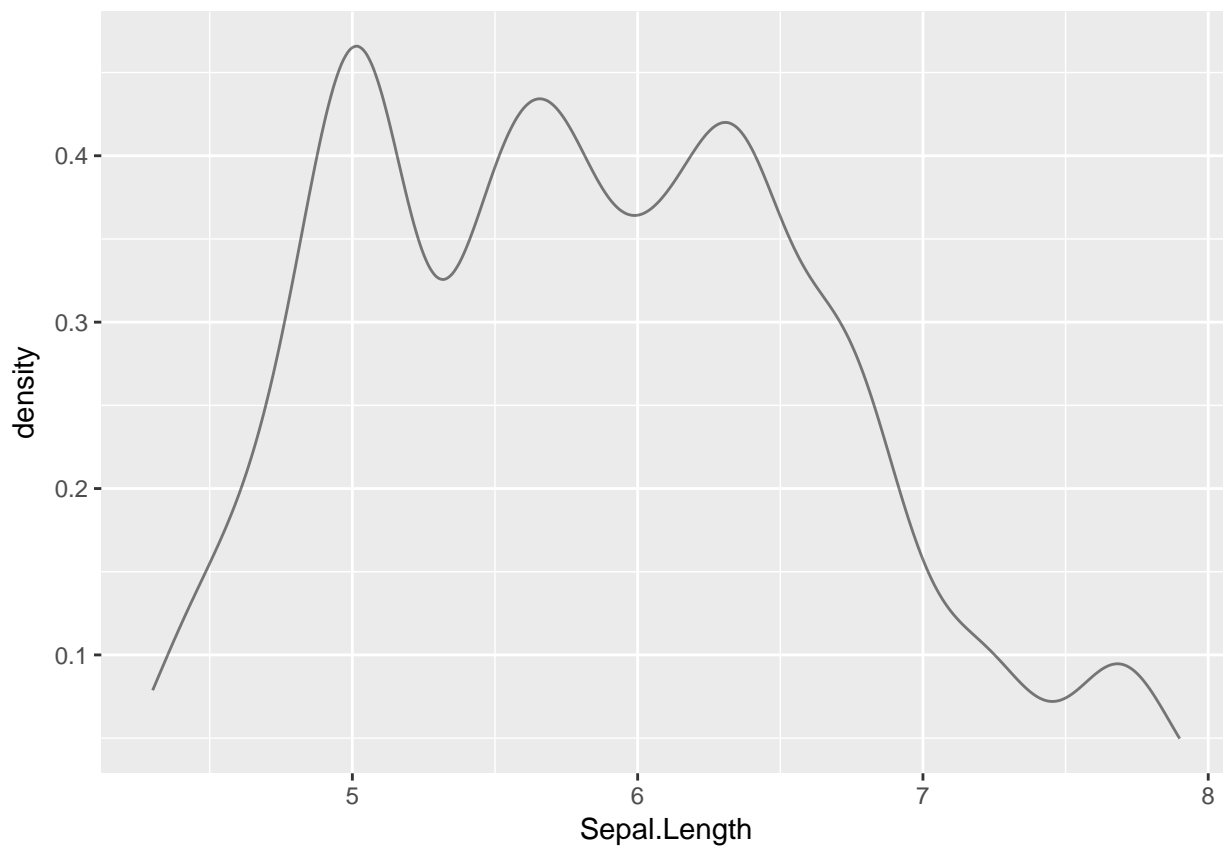
31	7.13270143	0.7519089987
32	7.74097597	-0.5816073956
33	9.13171716	-1.3348963632
34	9.50890512	-1.9364940241
35	7.48054469	0.6397971417
36	8.45649454	0.1623200988
37	8.98319882	-0.5541582898
38	8.56062490	-0.2686073922
39	7.43584562	0.8605892719
40	7.99599643	-0.0537378274
41	8.27215790	-0.5986282497
42	6.47967280	2.0742201695
43	7.65978151	0.4432510513
44	6.71143647	-1.0034513024
45	6.88724863	-1.0016002281
46	7.27470095	0.5905929779
47	8.18579890	-0.8100662308
48	7.58126010	0.4540714509
49	8.51144817	-0.7472727966
50	8.05232385	0.1103470642
51	-1.92650641	-0.3302436241
52	-2.21118264	-0.4976458405
53	-2.90655810	-0.1444036745
54	-2.21217081	1.7190884458
55	-2.82734979	0.3816148194
56	-2.76312977	0.9999554453
57	-2.96729844	-0.7291441115
58	-0.04584789	1.8044331516
59	-2.10126557	0.5657600008
60	-2.03772273	0.6940663687
61	-0.92008435	2.7620418494
62	-2.10978034	-0.1465326379
63	-1.08875553	2.3988056268
64	-3.07037581	0.5996380788
65	-0.41831193	0.1199178136
66	-1.53358826	-0.2553271700

67	-3.15331015	0.1898029330
68	-0.96653302	1.5013357513
69	-3.51041000	1.6565729007
70	-1.11604934	1.6162164791
71	-4.22077383	-0.7336497977
72	-1.11368738	0.4731599796
73	-4.11700948	1.3101939020
74	-2.65800143	1.2348852999
75	-1.50660553	0.3982435255
76	-1.73533016	-0.0128942406
77	-2.81199541	0.6503084931
78	-3.98048558	-0.2164369989
79	-2.90618229	0.2634167668
80	0.38014098	1.2736804538
81	-1.05972192	1.7803013707
82	-0.53948144	1.9152423883
83	-0.97473670	0.9180615649
84	-4.94070523	0.9375541590
85	-3.33285805	0.2573305712
86	-2.60851372	-0.9932178401
87	-2.56996737	-0.2335721119
88	-2.52625647	1.7623700443
89	-1.59669056	0.3029888924
90	-1.98823492	1.3017502252
91	-2.64637308	1.6197623217
92	-2.70033855	0.3126209306
93	-1.34477397	1.2050787130
94	-0.06804188	1.9793384428
95	-2.19066371	1.0073442612
96	-1.50281477	0.5608621666
97	-1.87695387	0.5562422215
98	-1.68615343	0.4657711637
99	0.75770475	1.0799032341
100	-1.73085250	0.6865632939
101	-8.68173998	-1.6302211183
102	-5.90676661	0.3652146310

103	-6.99269826	-0.4995161565
104	-6.26213639	0.3840865594
105	-7.53544381	-0.5885703350
106	-8.35031371	-0.1198989872
107	-4.86590985	1.0429171074
108	-7.17088209	0.5948846332
109	-6.86705099	1.2404037999
110	-7.79593990	-2.4817547834
111	-4.98068039	-1.1277667097
112	-5.88426154	0.3193277920
113	-6.22974285	-0.7116168505
114	-6.22457830	0.5246795775
115	-7.10565446	-0.9098997563
116	-6.37310645	-1.5771739811
117	-5.71255122	0.0295417731
118	-7.88503667	-1.9579567063
119	-10.00696181	0.4894796376
120	-4.98030448	2.1158407281
121	-6.95651395	-1.4326009253
122	-5.72037910	-0.1459119721
123	-8.48037381	0.5553125075
124	-4.67958706	0.2529885152
125	-6.49975159	-1.1471642866
126	-6.15054425	-0.2324029921
127	-4.39932375	-0.0002648138
128	-4.52323113	-0.3054911775
129	-7.07084386	-0.0808753156
130	-5.33399919	0.4548172636
131	-6.93910861	0.3998047933
132	-6.40693850	-1.8339503472
133	-7.33301502	-0.2941643710
134	-4.29724427	0.8408826468
135	-5.72890349	1.9307777502
136	-7.49453550	-0.9719811063
137	-7.27532362	-1.9389933245
138	-5.69035723	-0.1453635181

```
139 -4.35493577 -0.3500753962
140 -5.76993163 -1.0323978177
141 -7.25213165 -1.4480412701
142 -5.52006600 -1.6940200421
144 -7.56242654 -1.2421410306
145 -7.54843623 -2.0003205082
146 -6.06965117 -1.3394752557
147 -5.42376343 0.5353857182
148 -5.46268559 -0.6320804512
149 -6.58678778 -1.8486365257
150 -5.21891767 -0.0812674636
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	1
virginica	0	2	48



As you can see the LDA decision rule depends on x only through a linear combination of its elements. One drawback however of the LDA is the restrictive conditions associated with the algorithm. LDA assumes that features within the data set are a series of independent multivariate gaussians with identical covariance matrices. Additionally, LDA requires the number of features to be less than the sample size. As p approaches n , the performance of the model declines. If the assumption of uniform variance is highly off, then LDA can suffer high bias (Hastie, 153). On the iris data set, the LDA model had a missclassification rate of 0.0201342. Further exploratory data analysis shows that the variable sepal length does not satisfy the normality assumption. Such violations may become more apparent in a highly sparse setting.

Chapter 3

Results

Five models are trained on a random sample of 8,000 observations, and tested on a sample of 2,000 unseen responses. Both sets are comprised of indicator variables that denote the presence or absence of words in each response.

3.0.1 Run Time

The amount of time taken to learn the 8,000 responses in the training set varied substantially across all methods. The linear SVM and decision tree were amongst the fastest trained models taking just 43.3548701 and 31.3683829 seconds respectively. The multinomial logistic was the fastest model, taking just 48.2892339 seconds.

The higher runtime for the random forest model follows convention given that it is an aggregation of several learning methods. The runtime of the LDA model initially came as a surprise, however LDA computes discriminant scores $\delta_k(x)$ by finding a linear combination of the independent variables. In a highly sparse setting, this becomes a linear combination of over 10,000 variables for each of the 6 categories. The training time of the SVM, decision tree, and multinomial models were quite impressive, given the nature of the problem when applied to larger data sets.

model	accuracy	runtime
SVM	0.2000	43.35487
Multinomial	0.2250	31.36838
Decision Trees	0.6430	48.28923
Random Forests	0.2565	3279.40084
LDA	0.2610	6686.38202

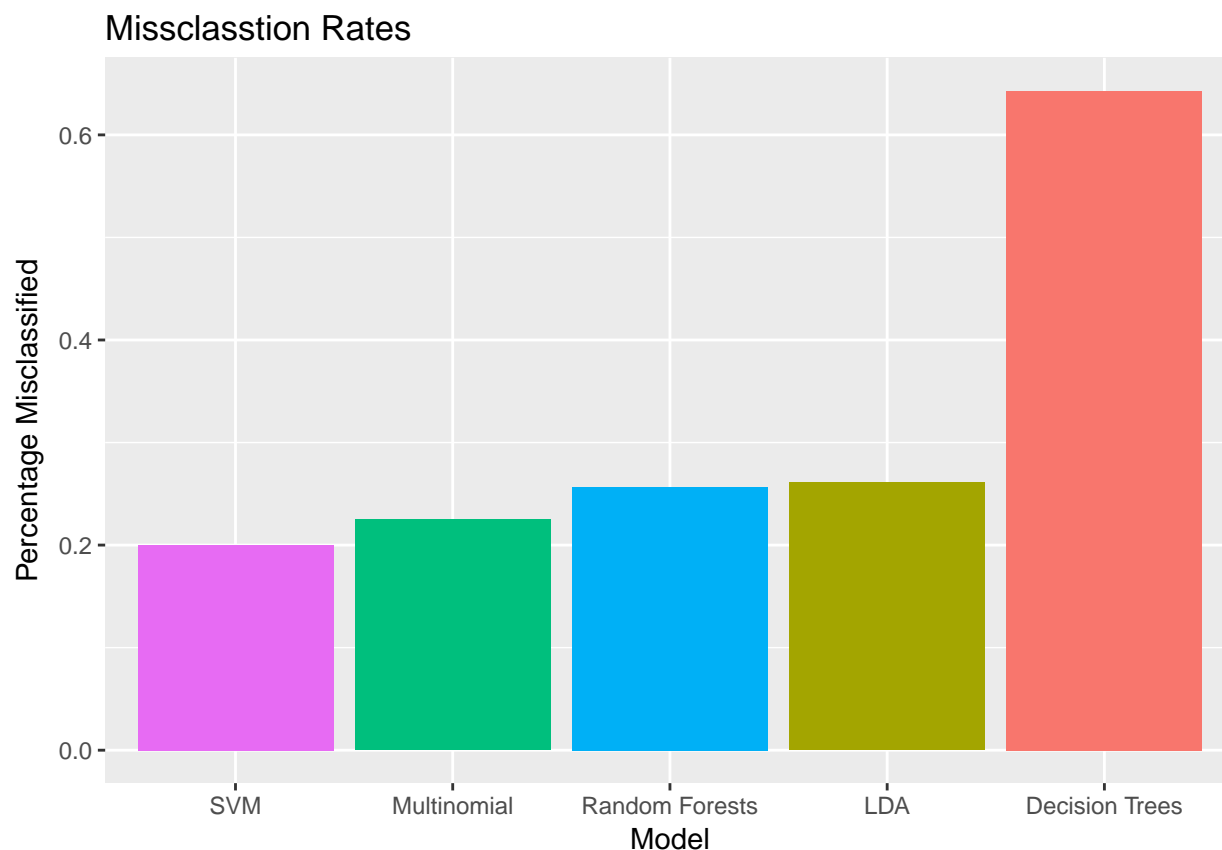


Figure 3.1: Bar Plot of Missclassification Rate

3.0.2 Classification Accuracy

As stated in the introduction the goal of this analysis was to compare classification error rates between models. SVMs had the lowest misclassification rate with 0.2. This was a slight improvement on the multinomial model (0.225 misclassification rate). LDA and Random Forest performed surprisingly well (0.261 for LDA and 0.2565 for rf). The decision tree produced the highest misclassification rate with 0.643 incorrectly categorized. Figure 3.1 shows the misclassification rates for all five algorithms tested.

While accuracy is a valid measure for information retrieval, in situations where classes are imbalanced, always predicting one class may yield a high accuracy. Precision and recall can be used to focus the evaluation on the correctly label categories (true positives). Precision is defined as the proportion of items i placed in the category correctly out of every algorithmic declaration of i , while recall is the proportion of items correctly guessed when it truly belonged in that category (Platt, 5). Low precision-high recall systems guess a label frequently, however a significant portion of labels are incorrect. High precision-low recall systems are the opposite, predicting a class less frequently but more accurately. Ideal systems have both high precision and

high recall, will return many results, with all results labeled correctly. The table below consists of precision-recall values for the category achievement for all 5 algorithms.

	SVM_PRECISION	SVM_RECALL
SVM	0.81	0.88
Multinomial	0.64	0.96
Decision Tree	0.49	0.95
Random Forests	0.72	0.89
LDA	0.75	0.78

The most efficient classifier (SVM) had high precision and high recall for the category achievement. Although the decision tree predicted achievement frequently, only 0.49 were correctly labeled. Interestingly enough, this was also this case for the multinomial model, which had a low precision of 0.64 and high recall of 0.96. However unlike the decision tree, the overall misclassification rate for the multinomial was quite low (0.225% labeled incorrectly). When their confusion matrices are compared, it is immediately apparent that two classes were not predicted in the decision tree model. Due to its greedy nature, the decision tree appeared to overfit on the noise in the training data. This leads to poor generalizability, and the algorithm only predicting the majority categories. On the other hand, the multinomial accurately predicted a high proportion of the other classes.

	achievement	affection	bonding	enjoy_the_moment	exercise	leisure	nature
achievement	625	242	34	208	13	110	31
affection	19	400	1	1	0	2	2
bonding	0	20	180	2	0	0	0
leisure	13	16	2	5	7	62	5

	achievement	affection	bonding	enjoy_the_moment	exercise	leisure
achievement	633	93	27	140	9	71
affection	16	567	5	13	0	14
bonding	1	7	183	1	0	0
enjoy_the_moment	5	5	1	58	1	2
exercise	0	0	0	0	9	0
leisure	2	6	1	4	1	87

	achievement	affection	bonding	enjoy_the_moment	exercise	leisure
nature	0	0	0	0	0	0

Conclusion

This experiment compares the performance of six machine learning algorithms tasked with classifying unstructured textual data. First, I discuss how the transformation of texts into document term matrices allows for words to be used as features in supervised methods. I also show how each of the algorithms creates a classifier based on these features, and predicts labels for new data. Due to issues with reproducibility, the bagging model discussed theoretically and was not evaluated further. My results demonstrate that accurate text classifiers can be trained from relatively simple models. Following Joachim's findings, it was empirically shown that SVM models consistently outperform competing methods in precision, recall. The high number of features and sparsity of the data aided in the creation of decision boundaries cast by the SVM by relying on the fact that the features are linearly separable. Although other methods such as Random Forests and LDA had high classification rates, their runtimes made them poor alternatives. A more detailed look at precision and recall for the LDA model highlights how its failed assumptions impacted the accuracy of the model. Likewise, the high variance of the decision tree is evident high error rate.

The two most important aspects of this experiment are the structure of the data, and the assumptions of the models. Further exploratory analysis may have shown why methods like LDA were not suitable for this data. Most of the limitations are a result of the high dimensionality of the data, or its incompatibility with the algorithms. The presence of the zero value in a sparse matrix provides no information while allocating memory for each 32-bit value. A document term matrix of the entire dataset occupies 25 megabytes of space. Using 15,000 observations, a more manageable 3.1 mb document term matrix is created. While this made computations feasible, reducing the size of the corpus and document term matrix reduces the number of unique words encountered in the training data.

For future studies, it would be advisable to test other natural language processing techniques. For example, how might multi-word (n-grams), noun phrases, and weighting affect the classification accuracy? Another interesting factor to consider is

the role of dimension reduction techniques such as principal component analysis and support vector decomposition. By projecting data onto lower dimensional subspaces, these methods may alleviate some of the issues related to sparse matrices, and provide improved algorithmic runtimes.

Appendix A

Appendix

In the main Rmd file:

```
library(acstats)
library(RJSONIO)
library(RCurl)
library(ngram)
library(tm)
library(wordcloud)
library(ggplot2)
library(glmnet)
library(text2vec)
library(data.table)
library(magrittr)
library(dplyr)
library(mosaic)
library(RTextTools)
library(e1071)
# library(caret)
library(tidyr)
library(stringr)
library(jsonlite)
library(tidytext)
library(rpart)
library(broom)
library(kableExtra)
```

```
#library(MASS) #lda
```

```
#Loading Data
```

```
setwd("~/Desktop/Statistics/Comps/Comps - Fayorsey/Comps-Fayorsey19E")
data <- read.csv("cleaned_hm.csv", stringsAsFactors = FALSE)

data$predicted_category <- as.factor(data$predicted_category)
```

```
#Creating Term Document Matrix
```

```
set.seed(7)
random_hm <- sample(1:nrow(data), 15000)
corpus <- Corpus(VectorSource(data$cleaned_hm[random_hm]))
skipWords <- function(x) removeWords(x,
  words = c(stopwords(kind = "en"), 'happy',
    'day', 'got', 'went', 'today', 'made', 'one',
    'two', 'time', 'last', 'first', 'going', 'getting',
    'took', 'found', 'lot', 'really', 'saw', 'see', 'month',
    'week', 'day', 'yesterday', 'year', 'ago', 'now', 'still',
    'since', 'something', 'great', 'good', 'long', 'thing',
    'toi', 'without', 'yesteri', '2s', 'toand', 'ing'))
funcs <- list(skipWords, stripWhitespace, removeNumbers,
  removePunctuation, tolower)
a <- tm_map(corpus, FUN = tm_reduce, tmFuns = funcs)
a_tdm <- TermDocumentMatrix(a)
m <- as.matrix(a_tdm)
v <- sort(rowSums(m), decreasing = TRUE)
d <- data.frame(word = names(v), freq = v)
d <- head(d, 10);d
```

```
#Displaying TDM for first three observations
```

```
a_dtm <- DocumentTermMatrix(a)
wa <- as.matrix(a_dtm)
```

```
wa[1:3,1:4]
```

#PieChart of Categories

```
ggplot(data, aes(x=predicted_category, fill=predicted_category))+
  geom_bar()+
  labs(title = "Distribution of Predicted Category")+
  guides(fill="none")+
  coord_polar(theta = "y", start=0)
```

#Summary of wordcount

```
count <- sapply(data$cleaned_hm, wordcount) # Counts number of words
summary(count)
```

#Distribution of Word Counts

```
category <- c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29", "30-34",
              "35-39",
              "40-44", "45-49", ">=50")
count_class <- cut(count, breaks = c(0,4,9,14,19,24,29,34,39,44,49,
                                     Inf),
                  labels = category, include.lowest = TRUE)
ggplot()+
  geom_bar(aes(x = count_class, fill = count_class))+
  ylim(0,30000)+
  labs(x = "Word Count", y = "Number of Happy Moments",
       title = "Word Count Distribution")+
  guides(fill = "none")
```

#TD-IDF

```
words <- data %>%
  unnest_tokens(word, cleaned_hm) %>%
  count(predicted_category, word, sort = TRUE) %>%
```

```

ungroup()

totalwords <- words %>%
  group_by(predicted_category) %>%
  summarize(total = sum(n))

words <- left_join(words, totalwords);words

tf <- words %>%
  bind_tf_idf(word, predicted_category, n);tf

tf %>%
  select(-total) %>%
  filter(n >=30) %>%
  arrange(desc(tf_idf));tf

```

#TD-IDF Graph

```

tf %>%
  arrange(desc(tf_idf)) %>%
  mutate(word = factor(word, levels = rev(unique(word)))) %>%
  group_by(predicted_category) %>%
  top_n(10) %>%
  ungroup %>%
  ggplot() +
  geom_col(aes(word, tf_idf, fill = predicted_category),
           show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~predicted_category, ncol = 3, scales = "free") +
  coord_flip()

```

#Creating containers for models

```

set.seed(57)

bin_data <- data[sample(1:nrow(data), 15000), ]

```

```
my.matrix1 <- create_matrix(bin_data$cleaned_hm, language="english",
                             removeNumbers=TRUE, stemWords=FALSE,
                             weighting=tm::weightTfIdf)

container1 <- create_container(my.matrix1, bin_data$predicted_category,
                              trainSize=1:8000, testSize =8001:10000,
                              virgin = FALSE)
```

#SVM model

```
set.seed(6)
start_svm <- Sys.time()
svm_model <- train_model(container1, "SVM")
end_svm <- Sys.time()

svm_results <- classify_model(container1, svm_model)
ac_svm <- mean(svm_results$SVM_LABEL != bin_data[8001:10000,9])

table(svm_results$SVM_LABEL, bin_data[8001:10000,9])

end_svm - start_svm

pr_svm <- create_precisionRecallSummary(container1, svm_results)
```

#Tree Model

```
set.seed(1)
start_tree <- Sys.time()
tree_model <- train_model(container1, "TREE")
end_tree <- Sys.time()

tree_results <- classify_model(container1, tree_model)
table(tree_results$TREE_LABEL, bin_data[8001:10000,9])

ac_tree <- (682+363+202+39)/2000
```

```
end_tree - start_tree
```

```
pr_tree <- create_precisionRecallSummary(container1, tree_results)
```

```
#Multinomial Model
```

```
set.seed(2)
```

```
start_multi <- Sys.time()
```

```
multinomial_model <- train_model(container1, "GLMNET",  
                                family="multinomial")
```

```
end_multi <- Sys.time()
```

```
multi_results <- classify_model(container1, multinomial_model)
```

```
table(multi_results$GLMNET_LABEL, bin_data[8001:10000,9])
```

```
ac_multi <- mean(multi_results$GLMNET_LABEL != bin_data[8001:10000,9])
```

```
end_multi - start_multi
```

```
pr_multi <- create_precisionRecallSummary(container1, multi_results)
```

```
#Bagging Model
```

```
# set.seed(3)
```

```
# start_bagging <- Sys.time()
```

```
# lda_model <- train_model(container1, "BAGGING")
```

```
# end_bagging <- Sys.time()
```

```
#
```

```
# bagging_results <- classify_model(container1, lda_model)
```

```
# table(bagging_results$BAGGING_LABEL, bin_data[8001:10000,9])
```

```
# ac_bagging <- mean(bagging_results$BAGGING_LABEL != bin_data[8001:10000,9])
```

```
#
```

```
# end_bagging - start_bagging
```

```
# pr_bag <- create_precisionRecallSummary(container1, bagging_results)
```

```
#Random Forests Model
```

```
set.seed(4)
```



```

start_rf <- Sys.time()
rf_model <- train_model(container1, "RF", ntree=10)
end_rf <- Sys.time()

rf_results <- classify_model(container1, rf_model)
table( rf_results$FORESTS_LABEL, bin_data[8001:10000,9])
ac_rf <- mean(rf_results$FORESTS_LABEL != bin_data[8001:10000,9])
end_rf-start_rf

pr_rf <- create_precisionRecallSummary(container1, rf_results)

```

#LDA Model

```

set.seed(5)
start_lda <- Sys.time()
lda_model <- train_model(container1, "SLDA")
end_lda <- Sys.time()

lda_results <- classify_model(container1, lda_model)
table(lda_results$SLDA_LABEL, bin_data[8001:10000,9])
ac_lda <- mean(lda_results$SLDA_LABEL != bin_data[8001:10000,9])
end_lda - start_lda

pr_lda <- create_precisionRecallSummary(container1, lda_results)

```

#Runtime table

```

time_svm <- end_svm - start_svm

# time_bagging <- end_bagging - start_bagging

time_rf <- end_rf - start_rf

time_multi <- end_multi - start_multi

time_lda <- end_lda - start_lda

```

#Decision Tree Confusion Matrix

```
table(tree_results$TREE_LABEL, bin_data[8001:10000,9])
```

#Multinomial Confusion Matrix

```
table(multi_results$GLMNET_LABEL, bin_data[8001:10000,9])
```


References

- Foster, D., et al. (2013). *Featurizing text: Converting text into predictors for regression analysis*. The Wharton School of the University of Pennsylvania.
- Hastie, T., et al. (2013). *An introduction to statistical learning with applications in R*. Springer.
- Joachims, T. (1994). *Text categorization with support vector machines: Learning with many relevant features* (pp. 1–6). University of Dortmund.
- Leopold, Edda, & Kinderman, J. (2002). Text categorization with support vector machines. how to represent texts in input space?, 1–20.
- Pang, Bo, & Lee, L. (2002). *Thumbs up? Sentiment classification using machine learning techniques* (pp. 1–20). Cornell University.
- Platt, et al, John. (1994). *Inductive learning algorithms and representations for text categorization*. Stanford University.
- Yogatama, D. (2015). *Sparse models of natural language text*. Carnegie Mellon University.