Biometric Monitoring and Bureaucratic Attendance in Jharkhand, India

Galen Murray *

Megan West †

March 19, 2016

*Department of Political Science, University of California, Los Angeles.

†Department of Political Science, University of California, Los Angeles.

Contents

1 Introduction and Motivation **1**

2 Data **2**

2.1 Data Collection and Background 2

2.2 Summary Statistics and Data Visualization 3

3 Model Selection and Results **6**

3.1 Logit (One-Versus-All) 7

3.2 LDA 10

3.3 QDA 12

3.4 Naive Bayes 14

3.5 LASSO 16

3.6 SVM 18

3.7 Random Forest 20

4 Discussion **22**

5 References **23**

1 Introduction and Motivation

In many developing countries, bureaucratic absenteeism (e.g. among teachers, doctors, etc.) is a problem (Chaudhury et. al 2006). One proposed solution is biometric attendance machines (BAM) that verify attendance via fingerprint scanners. While this technology reduces information asymmetries between principals (politicians) and agents (bureaucrats), leading research shows it is only a partial solution. Biometric machines were used to record health-worker attendance in Karnataka India in 2010. Despite increased technological monitoring, clinic staff were still absent 52% of the time and biometric machines had no effect on doctors (Dhaliwal and Hanna, 2014). Health worker absence in India remains pervasive.

In this project we explore the capabilities and limits of biometric attendance monitoring in the Indian state of Jharkhand. We do not yet have data on downstream outcomes to determine if increased attendance under biometric monitoring improves bureaucratic efficiency and service delivery. At the present time, this project is limited to exploratory analysis based on the limited data available. In particular, we are interested in predicting attendance based on the desirability of the department of employment, whether or not the individual holds a supervisory position, the month, and the weekday.

We compare the estimates and out-of-sample prediction estimates of six models: logit (one-versus-all), naive bayes, LDA, QDA, SVM, and random forest. We find that despite the powerful estimates of the one-versus-all logit model, we are unable to reliably predict attendance based on the aforementioned covariates. Furthermore, while some models, including QDA and random forest, are reasonably well-suited for this problem, logit, lasso, and SVM are poor methods for these predictions.

2 Data

2.1 Data Collection and Background

The Government of Jharkahand has installed biometric monitoring devices across the state which record the arrival and departure to the second of every employee in the office. This attendance data, along with employee metadata, is made freely available to the public.[1] On a monthly basis we scrape the attendance data from the GOJ website, collecting attendance records of about 5270 employees in approximately 68 departments in Jharkhand's civil service. Our primary output variable of interest is attendance, by employee, by day, which we code as a binary variable; **1 = Absent, 0 = Present**.[2] We predict daily bureaucratic attendance using *Month*, *Weekday*, *Supervisor* and *Important Departments*. *Supervisor* identifies whether or not the employee holds a supervisory role based on their job title description. Ideally we would have a mapping of the heads of each department or those in a supervisory role within their own department offices. Lacking that we code supervisors based on job title, the number of workers operating under the title in Jharkhand, and hand coding based on the details of the job description where available. *Important Department* is a binary for bureaucrats holding a position within a presitgious department.[3] *Month* and *Weekday* are categorical variables for the month and day of the week.

We have daily attendance data for 4 months for 5,268 employees, netting us a total sample of 323,400 daily-bureaucrat attendance observations. We divided the data into learning and test sets, randomly sampled by employee ID. The learning set includes 4270 employees, approximately 269,000 attendance measures. The test set includes 1,000 employees, approx-

[1]The data, attendance trends and reports are available at attendance.jharkhand.gov.in

[2]The scraped data contains the following covariates: Name, ID (employee ID), Status (employment status), Employment category (i.e. job type), Department, Date, Attendance status (which takes four values: Absent, Holiday, Leave, and Present), Time-in (time employee signed in for work), and Time-out (time employee signed out of work).

[3]We use the definition of important departments adopted by Iyer Lakshimi and Mani (2012)

imately 61,700 attendance measures. Finally, we exclude from the data set 11 days in which the biometric system registered fewer than 10 people as absent.

2.2 Summary Statistics and Data Visualization

Figure 1, provides the daily, average absence rate across all 5,270 employees for the months of June, July, August and October 2015.[4]. The sharp decreases in the absence rate in Figure 1, occur primarily at regular weekly intervals on saturday and are not indicative of the employee attendance pattern we are trying to predict.[5]

Figure 2 depicts daily average attendance rate after dropping days where absence is less than 15%. After cleaning the data, we see that June had a steady decrease in absences from roughly 40 percent to 30 percent. The absence rate then flattens out across July and August, with roughly 30 percent of employees marked absent, before spiking in October to over 50 percent for part of the month. [6]

[4]Figures are based on attendance for the training set

[5]While it is not uncommon for employees to work a half day on Saturday in India. In general, between 200-300 employees are present on Saturday in the training set, with very few absent, leading to the low attendance rates. In comparison, 2,950 employess are present, on average, between Monday and Friday, or roughly 70 percent.

[6]We ran all models discussed in the next section without dropping any attendance days and results were substantially similar.

Figure 1: Absence Rate

Figure 2: Absence Rate Above 15%

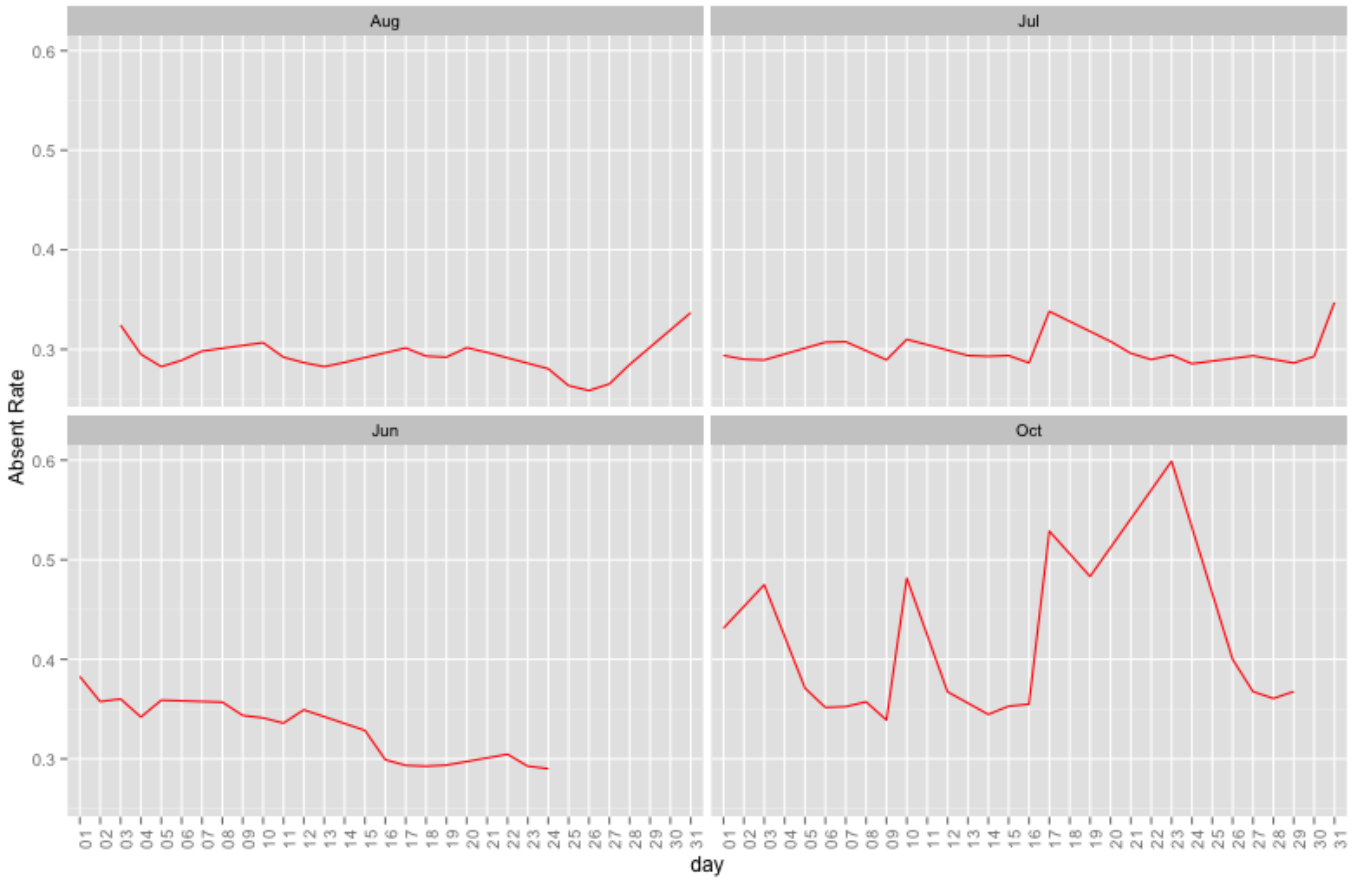


Table 1 provides summary statistics for absence and our predictors of interest for the entire dataset. On an average day, 30.2 % of employees are marked absent.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Absent	323,400	0.302	0.459	0	1
Important Dept.	323,400	0.197	0.398	0	1
Supervisor	316,332	0.151	0.358	0	1

3 Model Selection and Results

We compare the results and performance of seven models: logit, LDA, QDA, Naive Bayes, LASSO, SVM, and random forest. We focused our comparison on models that perform well in low-dimensional feature spaces, and thus excluded approaches such as neural networks. Because our covariates are all discrete and, with the exception of the month measurement, dichotomous variables, polynomial expansions do not dramatically change the models.

Below we summarize the results, prediction errors, and limitations of each of the above models, and Table 2 compares the prediction accuracy, mean squared error, ROC area under the curve, and predicted number of absences for each model. However, we find that the aforementioned features do not reliably predict bureaucratic attendance in any of the tested models. While, as discussed below, some of the predictive weakness may come from problems with each of the individual models, the consistent failure across models suggests that the position, desirability of a department, and month are weak predictors of attendance. Furthermore, the poor model performance can be specifically attributed to over-prediction of false positives. While most of the models have an overall prediction accuracy rate of approximately 70%, they all disproportionately under-estimate the number of absences.

Table 2: Model Accuracy

Model	Accuracy	MSE	Area	Pred. Absent (N)	Pred. Absent (%)
Logit	0.710	0.291	0.500	795	1.3%
LDA	0.709	0.291	0.504	795	2%
QDA	0.678	0.322	0.553	4,277	25.6%
Naive Bayes	0.682	0.318	0.548	3,654	21%
SVM	0.711	0.289	0.587	54	0.1%
Weighted SVM	0.510	0.451	0.587	54	0.1%
Random Forest	0.710	0.290	0.500	159	0.3%

3.1 Logit (One-Versus-All)

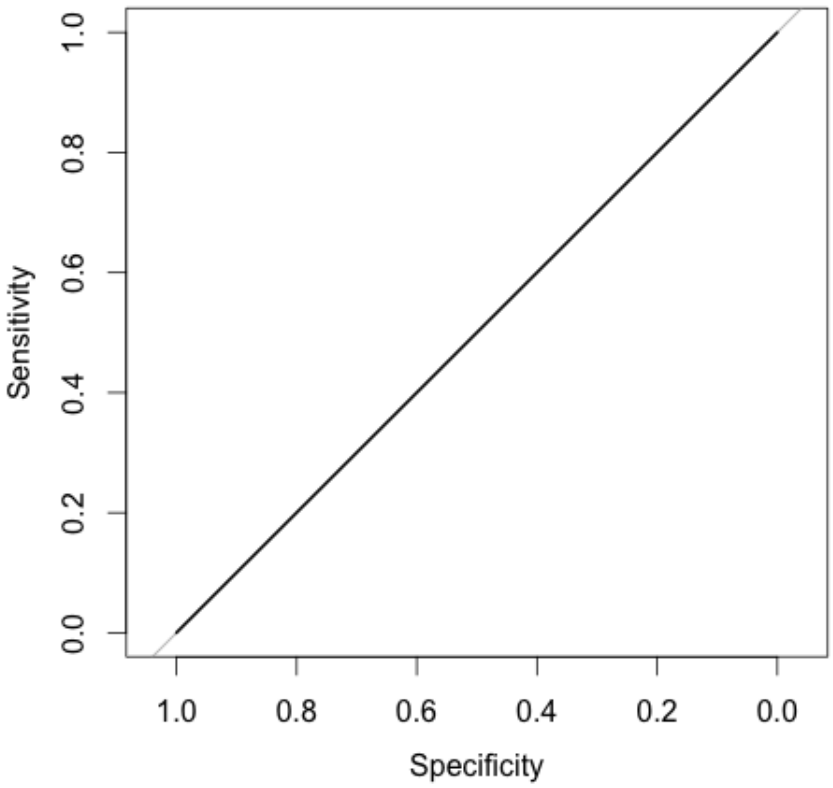
We initially tested the model using a standard logistic regression. Table 3 summarizes the model's estimated odds, and Figure 3 plots the ROC curve for the logit model. According to the logit model, individuals holding supervisory roles are approximately twice as likely to be absent as those who do not. Additionally, those who hold positions in important departments are approximately 15% less likely to be absent than those who do not. Features such as month and week day do not appear to significantly drive classification.

However, as summarized in Table 2, the model does not reliably predict absenteeism. Although it accurately classifies individuals in approximately 70% of cases, as the ROC curve indicates, it over-predicts attendance. In the test data set, approximately 28.9% of the total observations register as absences (16,737 out of 57,884). By comparison, the model predicts only 795 absences (1.3% of the total observations).

Table 3: Logit Model

	Dependent variable: Absent
	(1)
Supervisor	1.897***
	(0.013)
Important Dept.	0.821***
	(0.013)
Month: July	0.479***
	(0.012)
Month: August	1.015***
	(0.012)
Month: October	1.268***
	(0.016)
Weekday: Saturday	1.500***
	(0.060)
Weekday: Monday	1.012
	(0.014)
Weekday: Tuesday	0.926***
	(0.015)
Weekday: Wednesday	0.904***
	(0.014)
Weekday: Thursday	1.014***
	(0.014)
Constant	0.479***
	(0.012)
Observations	245,862
AIC	300,226

Figure 3: ROC Curve for Logit



3.2 LDA

Next, we use linear discriminant analysis to model the distribution of our 10 predictors seperately for each response class (*Absent, Present*) and invoke Bayes' rule to estimate $Pr(Y = k|X = x)$, where k is one of our two classes and x are our predictors of interest. Further LDA assumes that the density function of X ($Pr(X = x|Y = k)$) are Gaussian distributed and that the covariance matrices Σ_k are the same across both our classes *Absent and Present* .

Figure 4 plots histograms of the linear discriminant function values seperately for the Absent and Present groups. The histograms for both groups are nearly identical indicating that the training data does not seem to be easily seperable using a linear discriminant function. This is to be expected given our paucity of predictors and the likelihood of non-linearities in the data.[7]

Overall LDA correctly predicts 70.9% of bureacratic attendance. Results from the predictions made by the LDA classification are presented in Table 4. As the contingency table indicates, most of the error rate from the LDA model is due to underpredicting the number of bureaucrats who are absent. The model correctly classifies almost 99 percent of bureaucrats who are present. However, it only correctly classifies 2 percent of those who are absent (also see Figure 5 the Receiver Operating Characteristic curve, which is nearly flat indicating hardly any sensitivity i.e. correct prediction of absent bureaucrats). Given that we are interested in reducing bureaucratic absenteeism a 98% error rate for predicting the absent class does not help us identify features among our set of predictors that are indicative of bureaucrats who are unlikely to attend.

LDA likely does a poor job of classifying bureaucrats who are absent because it is trying to approximate the Bayes classifier, and give the lowest total error rate out of all classifiers. LDA

[7]Even when we include the department feature, a categorical variable for whether an employee belongs to 1 of 354 departments, our predictive power only slightly increases.

Figure 4: Linear Discriminant Function

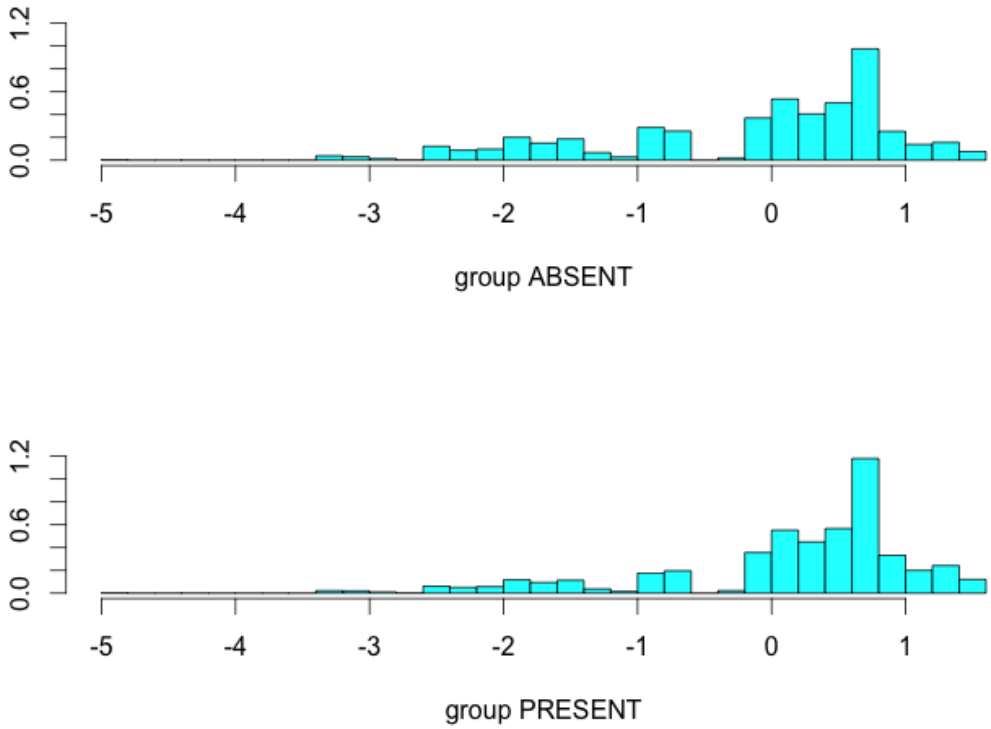
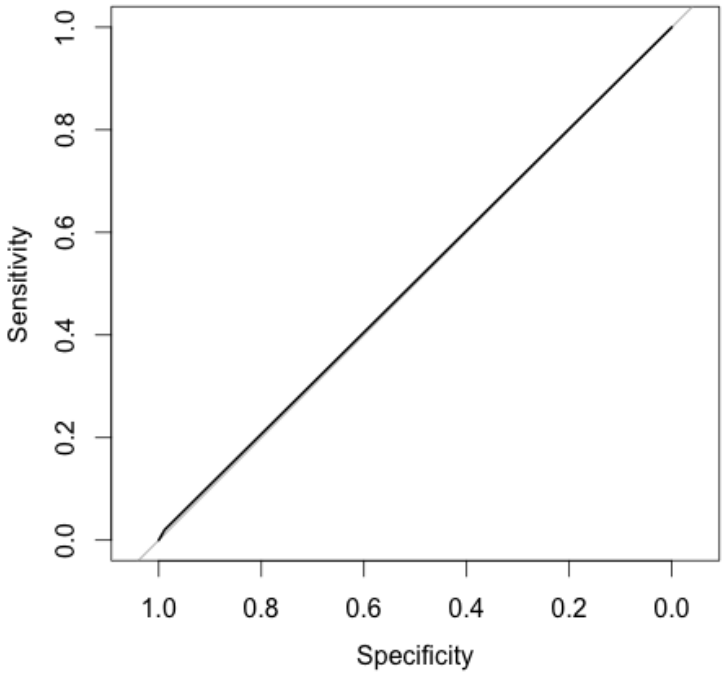


Table 4: LDA Confusion Table

	True	**Attend**
	ABSENT	PRESENT
Pred. ABSENT	332	463
Attend PRESENT	16405	40684

minimizes the total misclassification error "irrespective of which class the errors come from" (Gareth et al. ISLR Chapter 4.4). Finally, the LDA fitted model identified subordinates and bureaucrats from important/prestigious departments as more likely to be present.

Figure 5: LDA ROC



3.3 QDA

Quadratic discriminant analysis is similar to LDA but relaxes the assumption that the covariance matrices Σ_k are equal across the two groups.

Relative to LDA, QDA performs slightly worse with an overall error rate of 32%. However,

Table 5: QDA Confusion Table

	True	**Attend**
	ABSENT	PRESENT
Pred. ABSENT	4277	6137
Attend PRESENT	12460	35010

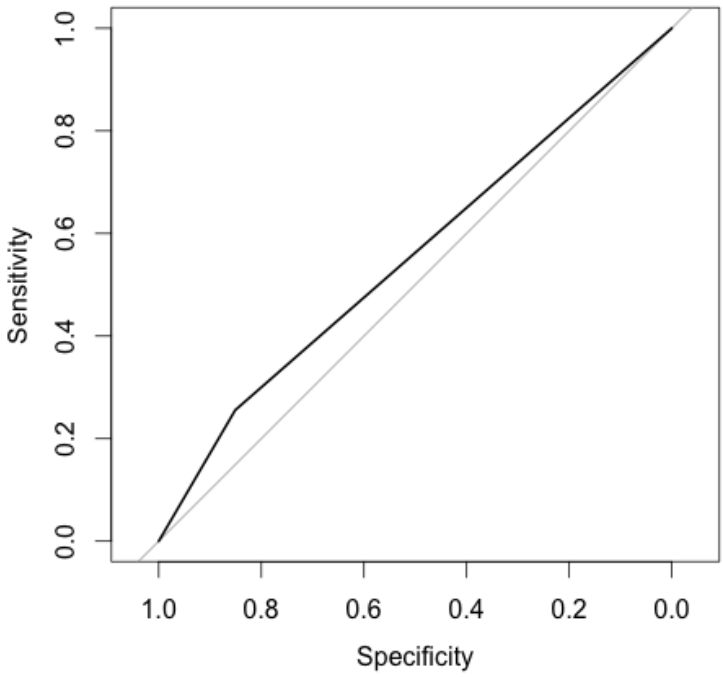
QDA does better in predicting which employees will be absent on a given day, with an error rate of 74 percent (see Figure 6 the ROC curve for QDA which indicates increased sensitivity). While this is a dramatic improvement over the LDA model which classified nearly every bureaucrat as *Present*, it is still an exteremely high error rate if we are interested in finding correlates of absence which could inform policies to reduce absenteeism. Second, the error rate for predicting bureaucrats who are present has fallen from 2% to nearly 15%. In other wrods, while the quadratic discriminant performs better at predicting which bureaucrats will be absent, it does so at the cost of a greater error rate in predicting employees who are present.

Interestingly, much like the results from the LDA analysis, bureaucrats in supervisory roles are more likely to be absent than those we categorize as holding subordinate positions. At the same time, bureacrats in prestigious or desirbable departments are more likely to attend. This could be read as bosses don't feel the same pressure to show up and those who have a desirable department want to maintain their higher status position. However, given the limited nature of this dataset and our imperfect coding of supervisory and important departments these correlations require further investigation.

Whereas LDA assumes that the covariance matrices are equal across our two classes, QDA does not make this restrictive assumption and is thus much more flexible. In our case, this seems like a reasonable tradeoff given that we only have 10 predictors which means estimating only an extra 90 parameters for the QDA model (120 parameters for QDA, 90 parametners for LDA). In other words, we are less concerned with reducing variance at the expense of greater bias. Given the large number of observations we have in our training set

(251,142) and the use of only 10 predictors it seems reasonable that QDA (the more flexible model) would outperform LDA in predicting which bureaucrats will be absent, albeit with a higher overall error rate. The slightly improved prediction rate in LDA probably speaks to the lack of informative features in our model. In fact, if you assigned everyone in the test set as *Present* based on the naive prior that 70 percent of employees are present, on aveage, in the trainsing set, then our models perform on par- or slightly worse- relative to an estimate that does not account for any predictors.

Figure 6: QDA ROC



3.4 Naive Bayes

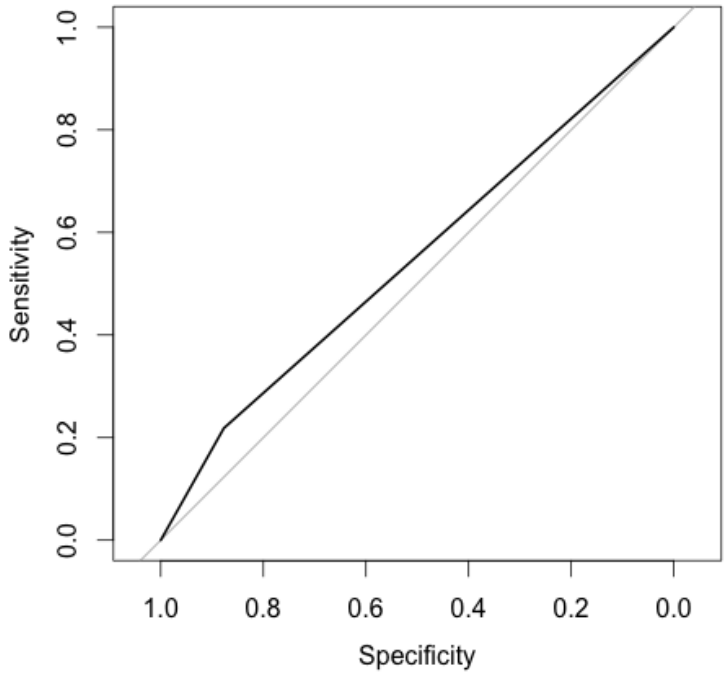
The naive bayes classifier assumes that our features are independent. Our naive bayes classifier performs similarly to QDA but, overall falls somewhere between the QDA and LDA models. The Naive Bayes classifier prediction power is on par with QDA, correctly

predicting 68 percent of bureacratic attendance in the test set. At the same time the Naive Bayes model has a higher error rate when predicting absentee bureaucrats (79%) but a lower error rate when predicting bureaucratic presence (12%). In this way, Naive Bayes falls somewhere between the QDA and LDA models in terms of senstivity and specificity (see Figure 7, ROC curve for naive bayes classifier).

Table 6: Naive Bayes Confusion Table

	True	**Attend**
	ABSENT	PRESENT
Pred. ABSENT	3654	5063
Attend PRESENT	13780	36916

Figure 7: Naive Bayes ROC



3.5 LASSO

The Fifth classification technique we explore is the LASSO. LASSO tends to shrink coefficient estimates towards zero, and at large enough values of the tuning parameter λ sets these coefficients exactly to zero. In essence, LASSO performs both shrinkage and variable selection, resulting in sparse models (Gareth et al. ISLR p. 219-220).

Figure 8 plots the coefficient estimates for various values of $log(\lambda)$.[8] As lambda increases the value of coefficents shrink towards zero. In fact, for a still, relatively small value of lambda ($exp(-3)$), all of the coefficient estimates are shrunk exactly to zero. This indicates that the optimal fit will only invovle a small amount of shrinkage and effectively reverts back to the least squares fit. Similarly, figure 9 plots the size of coefficient estimates as a function of the fraction of deviance explained (similar to r^2). Given the poor performance of our previous classifications techniques it is unsuprising that the LASSO model only explains a tiny percentage of the overall variance of bureaucratic attendance in the training dataset. While larger values of coefficients can explain more variation, the effect is quite small, suggesting that overfitting occurs towards the right hand side of the graph, where all 10 variables remain in the model and take on progessively larger coefficients.

We use cross validation (leave-one-out) in order to select the optimal value for λ, resulting in a $\lambda = 0.043$. As discussed above, this small value of lambda suggests that we are effectively reproducing the least squares estimate. The final model is extremely sparse (see table 7) and delivers. In fact the only variable not set to zero is *Supervisor* which has a large, positive coefficient indicating that supervisors are more likely to be absent than subordinates. This concurs with the other classification techniques we explore. However, unlike other models, LASSO does not attribute any predictive power to the important department variable.

[8]The numbers across the top of the plot indicate the number of coefficents not set to zero in the model for a given value of λ.

Figure 8: Estimated Coefs from LASSO while varying λ

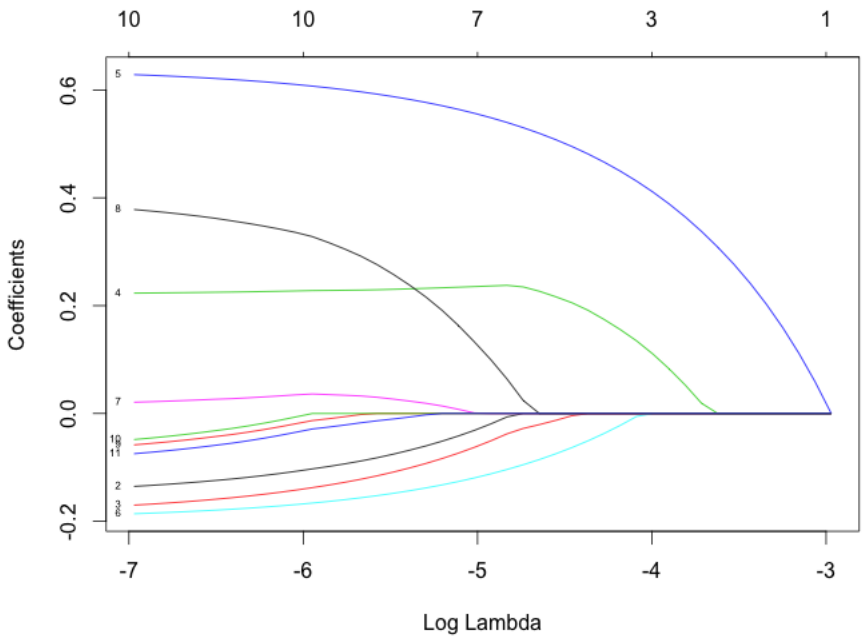


Figure 9: Estimated Coefs from LASSO vs. Explainded Deviance

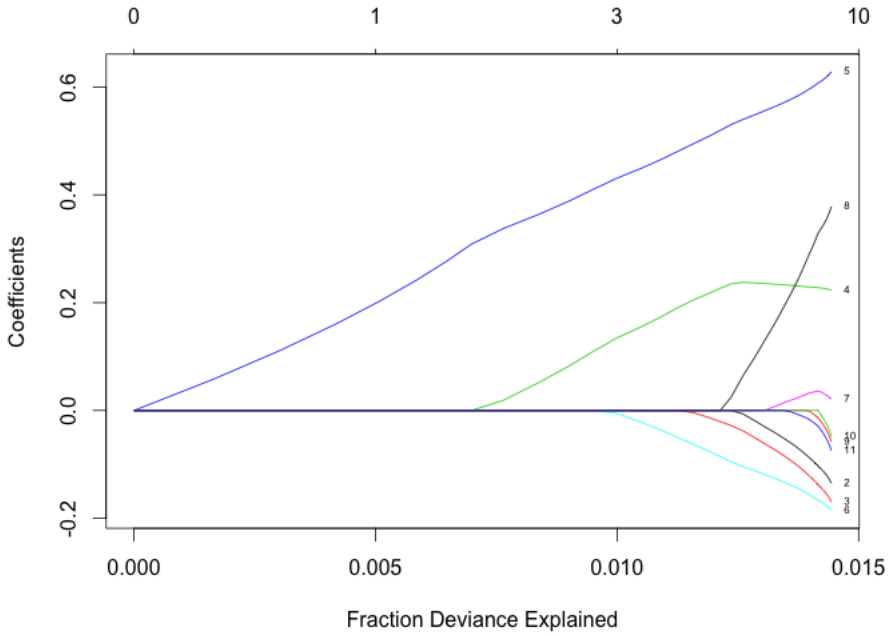


Table 7: LASSO Estimated Coeffcients, $\lambda = 0.043$

	1
Intercept	-0.815
Intercept	0
July	0
June	0
October	0
Supervisor	0.112
Important	0
Monday	0
Saturday	0
Thursday	0
Tuesday	0
Wednesday	0

3.6 SVM

The sixth classification technique we used was SVM. SVM seeks to maximize the margin between the classes of interest. It identifies observations that rest at the edge of the margin, creating support vectors, and predicts the classification of each observation based on where each falls relative to each support vector. Because very large data sets often lead to large support vectors, SVM requires more processing time for larger data sets and typically performs better with relatively fewer observations. Therefore, we randomly sampled 50,000 observations for each the training and test sets from the original training and test sets. Additionally, we test two SVM models: one weighted to account for the under-representation of absences, and one unweighted as a comparison.

Figure 10 plots the ROC curves for each model, and Table 8 summarizes the out of sample mean squared errors and area under the ROC curves for the weighted and unweighted models.

While the unweighed SVM did reduce the incidence of under-predicting absence, neither the weighted nor unweighted SVM improved the overall prediction error compared to the other classification models. The unweighted model performed comparably to other classifica-

Figure 10: ROC Curves for SVM

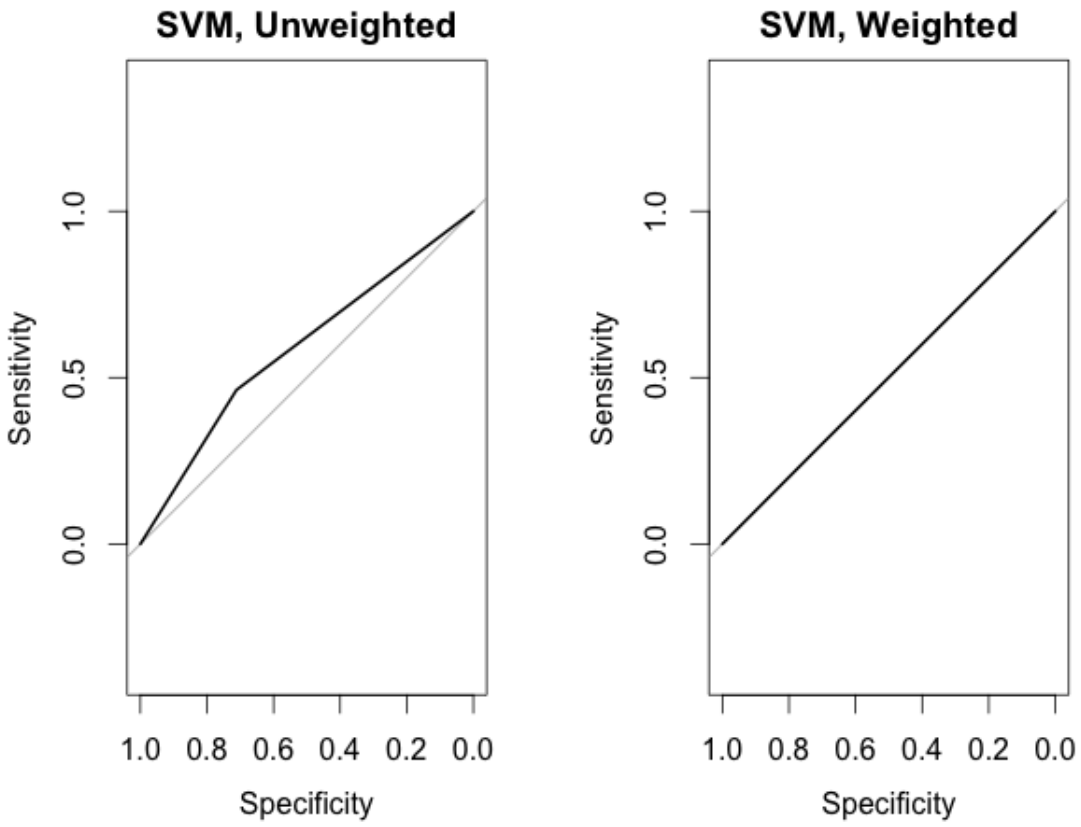


Table 8: Model Accuracy: SVM

Model	Accuracy	MSE	Area	Pred. Absent (N)	Pred. Absent (%)
Unweighted SVM	0.711	0.289	0.587	54	0.1%
Weighted SVM	0.510	0.491	0.501	24267	48.5%

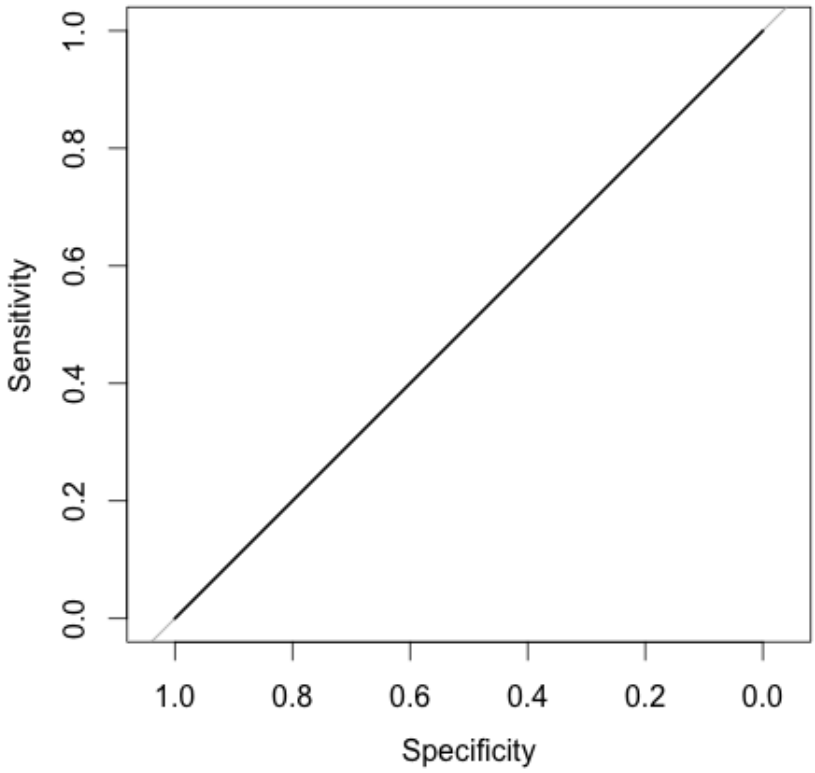
tion techniques, accurately classifying approximately 71% of the out-of-sample observations and under-predicting absence. The weighted model's prediction performs only slightly better than an at-random classification assignment, only accurately predicting 51% of the observations, and over-predicting absence. The poor performance is likely due to the fact that the features do not adequately separate the observations. Out of 50,000 observations, the unweighted model had 30,878 support vectors while the weighted model had 45,496 support vectors, indicating that the overwhelming majority of the observations lie at the margin rather than in a well-defined space. The weighted model was balanced to place more emphasis on the absences, and therefore was less prone to the under-prediction error present in the other models. However, as the features do not appear to inform the model, this did not improve the estimation, and the model classified about 50% of the observations as absent.

3.7 Random Forest

The final classification method we tested was Random Forest. Random forest is an extension of decision tree classification that uses bootstrapped sampling of the training dataset and random noise in the model selection process in order to reduce the problem of over-fitting inherent in decision tree models. Random forest begins by bootstrap sampling the training set and, for each new set, creates a typical decision tree clustering the observations based on the predictors, beginning with the predictors that explain the greatest degree of variance, with some randomness added to the attribute selection at each node and within each tree. The decision trees are then aggregated to create one set of predicted classifications.

Figure 11 presents the ROC curve for the Random Forest model and Table 9 summarizes

Figure 11: ROC Curve for Random Forest



the prediction accuracy.

Table 9: Model Accuracy: Random Forest

Model	Accuracy	MSE	Area	Pred. Absent (N)	Pred. Absent (%)
Random Forest	0.710	0.290	0.500	159	0.3%

Random forest did not add any leverage to the classification, performing as poorly as the previously tested models. Like the other classification models, it greatly under-predicted absenteeism, predicting only 159 absences out of 50,000 observations. As with the other models, we expect this indicates that the features do not inform classification. Random forests branch and form new nodes in order of variable information: the most informative

features will form the earlier nodes. Therefore, the models tend to perform well when a small number of features explain the majority of the variance and perform less well when the models are noisy or the classes are not well-defined by the features.

4 Discussion

Our investigation finds there is little evidence that the preliminary covariates predict attendance. While some models marginally improve the out-of-sample fit, overall every model substantially under-predicted absenteeism and produced poor out-of-sample classification predictions. We conclude that the most likely cause is a weak relationship between the covariates and attendance. While most models predict a significant relationship between the supervisor and important department features, the relationship between these features and the predicted classification was inconsistent across models. The most robust finding was that supervisors are less likely to attend than subordinates. On the other hand features such as *Important Department* are less consistent. For example, the logit model predicts that employees of important departments are more likely to attend than employees of departments deemed less important. However, the LASSO model shrinks the coefficient estimate for *Important Departments* to 0.

It is possible that a richer set of features would provide a more precise classification model. Other features worthy of exploration include the geographic location of a department, or distance of the department office to a major city and the size of the department, as well as features specific to the employee. At this time, however, the Government of Jharkhand has limited the amount of information publicly available on the employees.

5 References

1. Chaudhury, Nazmul, et al. "Missing in action: teacher and health worker absence in developing countries." The Journal of Economic Perspectives 20.1 (2006): 91-116.

2. Dhaliwal, Iqbal, and Rema Hanna. Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India. No. w20482. National Bureau of Economic Research, 2014.

3. James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

4. Iyer, Lakshmi, and Anandi Mani. "Traveling agents: political change and bureaucratic turnover in India." Review of Economics and Statistics 94.3 (2012): 723-739.