This chapter summarizes our first round of validation testing for the R version of the Paid Leave Microsimulation model.

## X.1 Introduction

The FMLA survey has data on leave taking that our simulation model wishes to impute into an ACS data set based on a selection of explanatory characteristics and imputation method. But how do we know what method to use? Accurate imputation of variables is paramount for ensuring the simulation model produces valid estimates. Imputation is in essence a prediction of an unobserved characteristic. To test the validity of imputation method, we use each method to “predict” the characteristic of known data and then evaluate how well the predictions did.

We are testing each method’s ability to accurately impute leave taking behavior in a number of different ways. The imputation of variables will always carry some risk of error, so our model only imputes from FMLA to ACS necessary variables which are not found in ACS.

There are many different methods by which we could conduct these imputations, and many different ways. For this testing memo, we have completed the coding of 6 different methods for imputing these variables into the ACS:

1. Logit Regression (denoted logit in the model code)
2. Random Forest Classifier (random forest)
3. Naïve Bayes Classifier (Naïve Bayes)
4. Ridge Regression Classification (ridge class)
5. K=1 Nearest Neighbors (KNN1)
6. K=5 Nearest Neighbor, based on majority voting (KNN multi)

We have implemented a 7th method, random draw imputation (random), which we use to compare each method to. For each observation in the test data set, the random draw method will simply pick a random observation in the training data set, and assign that training observation’s value to be the testing observation’ value for the imputed variable. All 7 methods uses the same parameters other than the imputation method parameter.

We present a comparison of the different imputation methods in three different ways:

1. FMLA-to-FMLA performance results – individual-level leave taking
2. FMLA-to-FMLA performance results – population aggregate leave taking
3. FMLA-to-ACS performance results – predicted vs actual benefits paid out for California, Rhode Island, New Jersey

FMLA to FMLA performance results are performance results resulting from splitting the FMLA data set into two parts randomly; a training set and a testing set. The training set is used to calibrate the predictive model with every imputation methods. This model is then used to predict leave taking in the test data set, and we evaluate the performance of these predictions against the actual leave taking values in the testing data. We then calculate performance measures such as accuracy, recall, and performance. We test both the predictive performance of population aggregate estimates (how many leave takers are there?) and individual-level estimates (who takes leave?). This distinction is illustrated by Exhibit X below; in this example, method 1 performs better at predicting aggregate levels of leave taking, while method 2 performs better at predicting who takes leave at an individual level. We want to know how each method will stack up with one another in both of these dimensions.

**Exhibit 1.**

**Aggregate versus Individual-Level Performance**

A, C, and D are Leave Takers – 66% accuracy

A and C are Leave Takers – 50% accuracy

|  |  |  |
| --- | --- | --- |
| ***Population-Level Aggregate Error*** | | |
| *Actual*  Non- Leave taker  2 Leave Takers  Leave  taker | *Predicted - Method 1*  2 Leave Takers – No Error | *Predicted - Method 2*  3 Leaves Takers – 50% overestimate |
| ***Individual-Level Error*** | | |
| *Actual*  A and D are Leave Takers | *Predicted - Method 1* | *Predicted - Method 2* |

Because ACS does not have data on actual leave taking behavior, we cannot validate the predictive performance of using the measures above. But we do have data on actual benefits paid out by states with PFL programs (CA, RI, and NJ). We used this data to test our model’s ability to predict benefits outlayed by each state program. For each state, we simulated leave taking and program participation in the state’s ACS population. We used model parameters that mirrored the real-world rules and restrictions of the respective state PFL program (the full specification of these parameters are included in Appendix X). A valid microsimulation model should produce estimates for benefits outlayed similar to the amount of benefits actual PFL programs paid out.

For this test, we used 2012-2016 5-year ACS data. This data was selected to minimize time distance from the administration of the FMLA survey in April 2011-2012. The farther we get from the FMLA survey predicting leave taking behavior, the more likely temporal variation in actual leave taking behavior is to contaminate the observed error between model estimates against actual state program data. To smooth out any randomness introduced by stochastic components of the imputation methods, all results presented are the mean result of 100 simulations. We also examined the variance due to these stochastic components for a single simulation, and found it to be minimal. As a result, all error bars represent solely the 95% confidence interval of estimates due to sampling error based on the FMLA survey replicate weights.

**X.1.1 Selection of Imputed Variables to Use for Testing**

We selected five different variables to test imputation performance:

1. Program Benefits Outlayed
2. Leave Takers
3. Number of Leaves Taken
4. Leave Needers
5. Proportion of Pay Received from Employer while on Leave (Prop\_pay)

These variables were selected for their importance and model sensitivity. We briefly justify the selection of each of these variables below.

***Program Benefits Outlayed.*** One of the main policy questions this simulation model will be able to answer is “how much benefits should a state paid leave program expect to outlay annually?” Because benefit outlays are not present in the FMLA survey, we cannot perform an intra-FMLA performance testing. However, a well-calibrated model should be able to estimate annual outlays of current state programs. This is accomplished by running the simulation from their state’s ACS population, selecting a simulated leave program that replicates the rules and restrictions of the actual state program. There are three states with sufficient historical data on benefit outlays to perform this test on: California, Rhode Island, and New Jersey.

***Leave Takers/Number of Leaves Taken.*** Can the model accurately predict how many individuals in the simulated population will take leave? This is an important part of calculating simulated benefit outlays. This benefit outlay can be generally thought of as the multiplication of a few different terms:

So, the model’s ability to predict level of leave taking in a population is related to the model’s ability to predict benefit outlays.

***Leave Needers.*** In the simulation, financially sensitive leave needers are assumed to take leave in the presence of a state leave program. So, our model needs to accurately estimate leave needing in order to be properly estimate leave taking in the presence of a state leave program.

***Proportion of pay received from the employer*** is a key variable to impute for determining program participation. Simulated individuals will choose to participate in the program if they would receive more benefits from the program than they would their employer. As a result, accurate program participation estimates rely on accurate estimates of employer leave pay. So, we examined how well each model does at predicting this proportion. While proportion of pay might seem to be a continuous variable, it is actually discrete in our model. This is because the FMLA survey is asked respondents to identify in which of seven ranges the proportion of pay received fell into: “None”, “One quarter or less”, “More than one-quarter but less than half”, “About half”, “More than half but less than three-quarters”, “Three quarters or more”, or “Full pay”. The model assigns the mid-point of each of these ranges to each individual. So this variable remains a numeric value, but can only take on one of seven values.

## X.2 Results

We interpret and discuss the results of our validation testing in this section. Our discussion is reflective of the primary goal of the model: accurately predicting the benefits outlay of a simulated leave program for any given state. Below is a summary of the preliminary conclusions we have drawn from our testing so far.

**Preliminary Conclusions**

* **KNN1 is the strongest performing method so far.** We plan to use KNN1 as the default imputation method for the model should these results hold through further testing.
* **No model’s individual-level performance is consistently better than random draws.** Rankings of individual-level performance amongst methods are not consistent either.
* **Better individual-level predictive performance tends to not be related to overall population-level performance.** Given our research question ultimately care most about an accurate population-level performance (and the extent to which individual-level performance affects population-level performance on other data sets is unknown), we place more weight on the methods that perform best with overall population-level predictions. This is validated by a comparison to actual benefit outlay results; models tend to perform similarly to their FMLA population-level performance results.
* **Together, this suggests the limiting of the use of the microsimulation model to answer research questions with aggregate-level results**. For example, the model would be more confident in answering “how many individuals will take leave in State X?” than “which individuals will take leave in State X”.

Below, Exhibit X is a summary of the comparative scores for each test that we conducted on the model. It is an approximate measure of how well each method did comparatively to other methods on each measure. While the weight to put on each performance measure is not necessarily equivalent, we see this table as a good visual aid to generally compare method performance across all the different measures we tested. We provide further discussion and speculation on what measures should be considered more or less important in the remainder of the section.

Broadly speaking, the more “Green” boxes in each column, the better the column’s method has performed. Conversely, the more “Red” boxes, the worse the performance. From the general patterns observed in this exhibit, our preliminary conclusions are:

* KNN1 is the best-performing method,
* Naïve Bayes is the worst-performing method,
* Ridge Class, Random Forest, KNN Multi and Logit are somewhere in the middle of those two.

**Exhibit 2.**

**Method Performance Summary**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exhibit # | Measure | Measure Type | Random | Logit | KNN1 | KNN Multi | Random Forest | Naïve Bayes | Ridge Class |
| 3 | Benefits Outlayed – RI | FMLA-to-ACS |  |  |  |  |  |  |  |
| 4 | Benefits Outlayed – NJ | FMLA-to-ACS |  |  |  |  |  |  |  |
| 5 | Benefits Outlayed – CA | FMLA-to-ACS |  |  |  |  |  |  |  |
| 6 | Predicted/Actual Leave Takers | FMLA-to-FMLA Aggregate | N/A |  |  |  |  |  |  |
| 7 | Predicted/Actual Number of Leaves Taken | FMLA-to-FMLA Aggregate | N/A |  |  |  |  |  |  |
| 8 | Predicted/Actual Prop Pay | FMLA-to-FMLA Aggregate | N/A |  |  |  |  |  |  |
| 9 | Predicted/Actual Leave Needers | FMLA-to-FMLA Aggregate | N/A |  |  |  |  |  |  |
| 10 | Leave Takers Accuracy | FMLA-to-FMLA Individual |  |  |  |  |  |  |  |
| 11 | Prop Pay Accuracy | FMLA-to-FMLA Individual |  |  |  |  |  |  |  |
| 12 | Leave Needers Accuracy | FMLA-to-FMLA Individual |  |  |  |  |  |  |  |

**Legend**

Poor Performance Good Performance



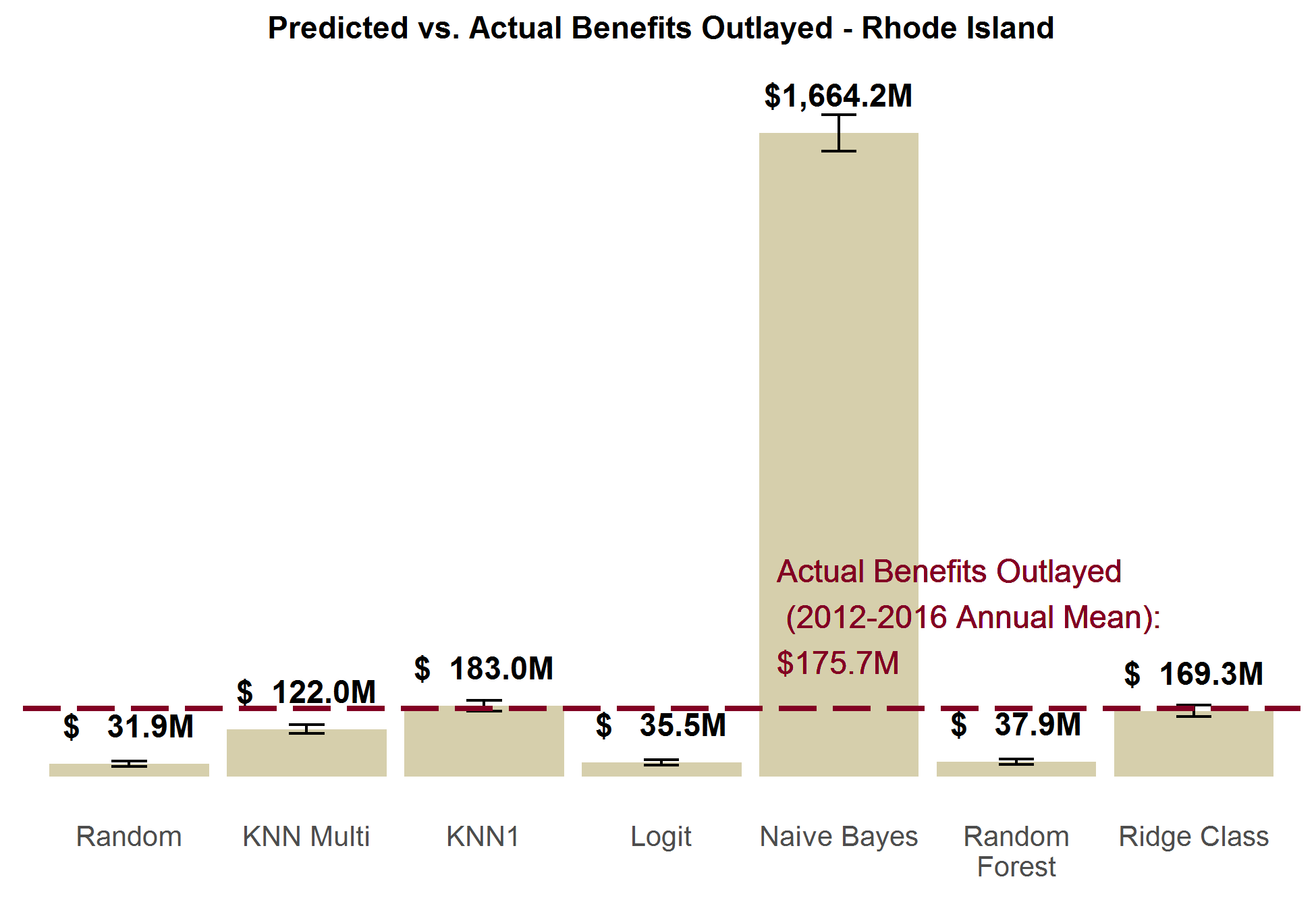
### In the next three subsections, we present and discuss the results of each of the three measure types (FMLA-to-ACS, FMLA-to-FMLA Aggregate, and FMLA-to-FMLA Individual).

### X.2.1 Benefit Outlays

There are three states with sufficient historical data on benefit outlays to perform this test on: California, Rhode Island, and New Jersey. Results from these simulations are shown below.

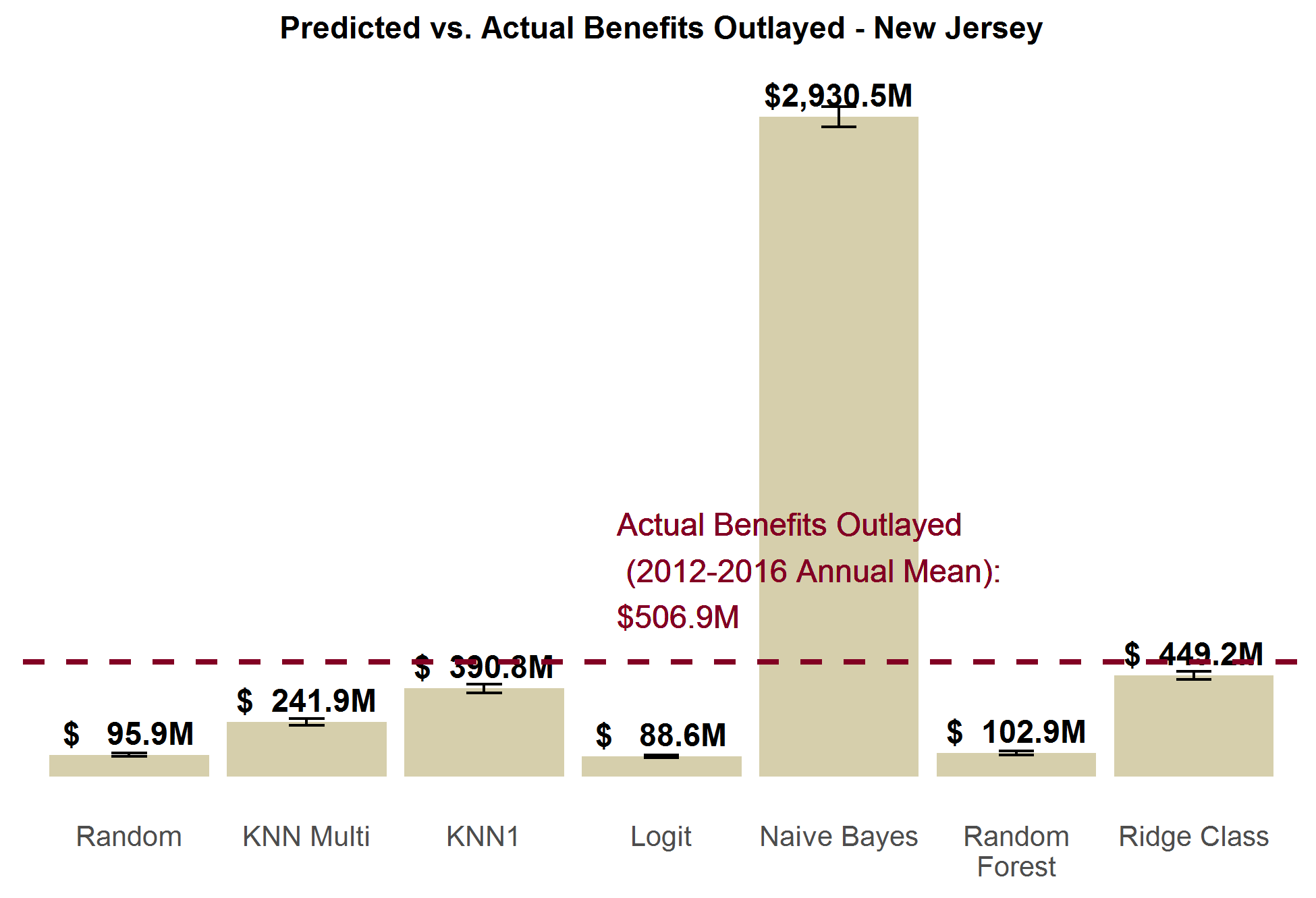
***Rhode Island.*** In Rhode Island, the actual program paid out approximately $176 million in benefits annually between 2012 and 2016. As seen in Exhibit X below, two of the prediction methods (KNN1 and ridge class) successfully capture the actual value within their 95% estimate confidence interval. KNN multi performs next best, but still under-predicts benefits by $53 million. Logit and random forest methods even more drastically under-predict benefit outlays, only estimating approximately $36 million and $38 million respectively. These results are similar to random draws, which we yield an estimate of $32 million. Naïve Bayes exorbitantly over-predicts benefit payouts by almost ten-fold, with an estimate of $1.6 billion.

**Exhibit 3.**

****

***New Jersey.*** New Jersey paid out an average of $507 million in benefits annually from 2012 to 2016. We see similar patterns in the predicted benefits outlayed as in Rhode Island. KNN1 and ridge class come closest to correctly predicting benefits. But in the case of New Jersey, both methods underestimate benefits by a statistically significant margin. Similarly, logit, random forest, and KNN multi methods all more drastically undershoot benefit estimates, while Naïve Bayes drastically overshoots it.

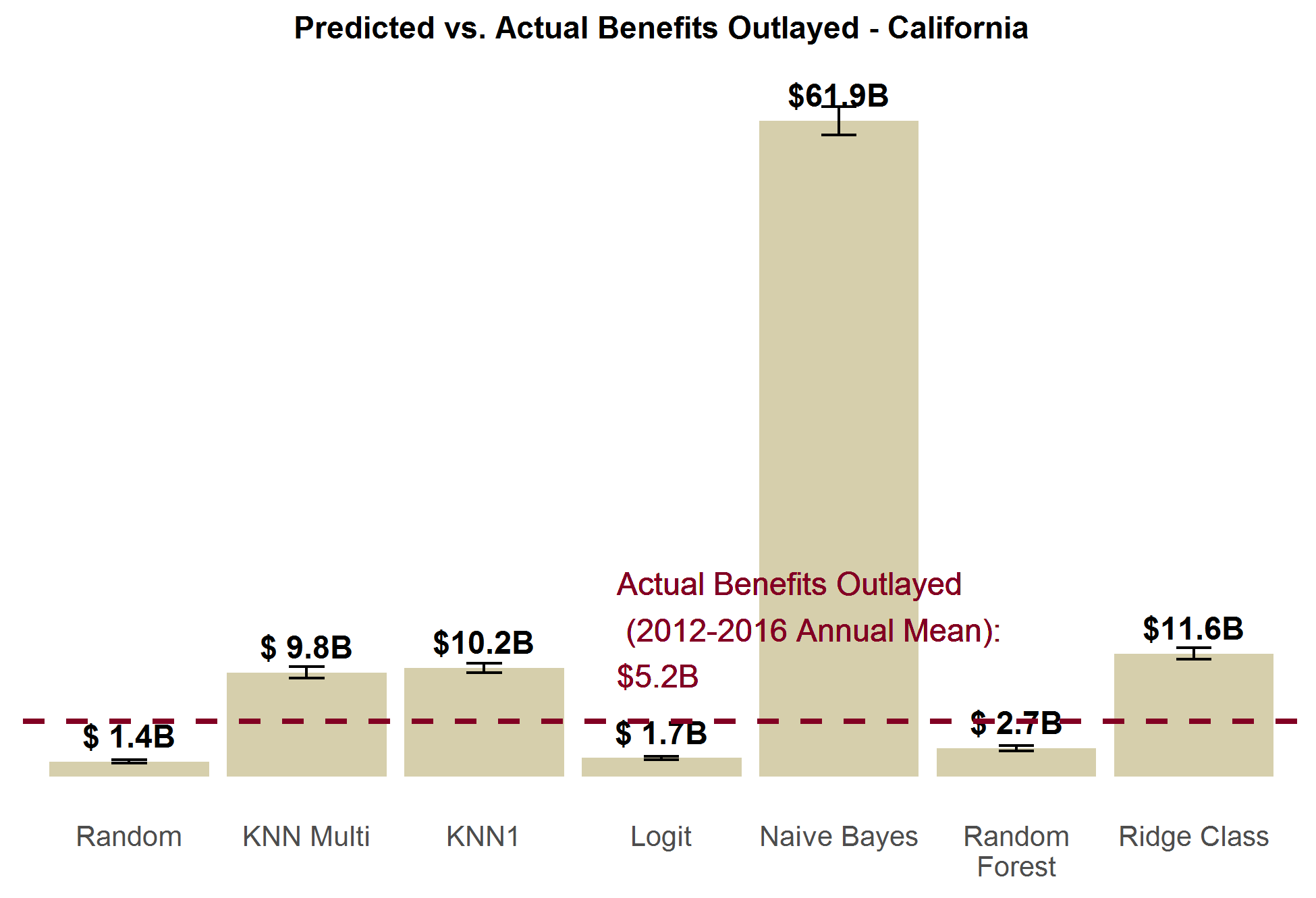
**Exhibit 4.**



***California.*** Of the current models, random forest comes closest to estimating California’s actual benefit outlay, though it still arrives at an estimate that about half of the actual outlays ($2.7 billion vs. $5.2 billion). Conversely, KNN multi, KNN1, and ridge class all overestimate benefits outlays by twice the actual benefits paid out. Like the in other states, Naïve Bayes massively overstates the benefits by over tenfold, and the random / logit methods drastically understate the benefit outlays.

We are investigating reasons why California’s predictions are somewhat worse than the other states overall.

**Exhibit 5.**

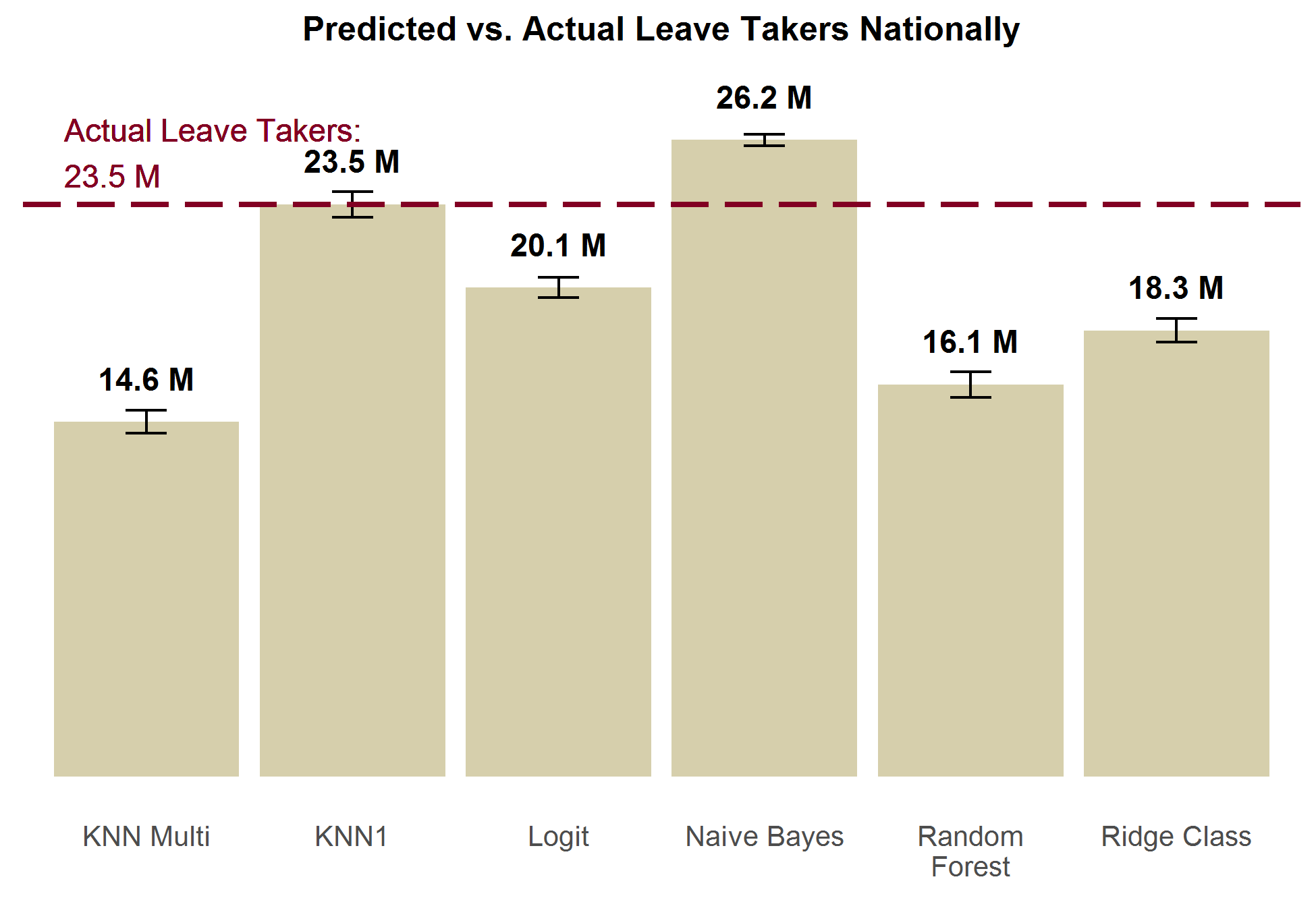


### X.2.2 FMLA-to-FMLA Performance in Aggregate

When we use part of the FMLA data to make predictions on the other part of FMLA, we can see how well the methods perform at predicting population levels of leave taking. We test the aggregate performance of a number of different variables imputed by the model: number of leaver takers, number of leaves taken, proportion of pay received from the employer, and number of leave needers.

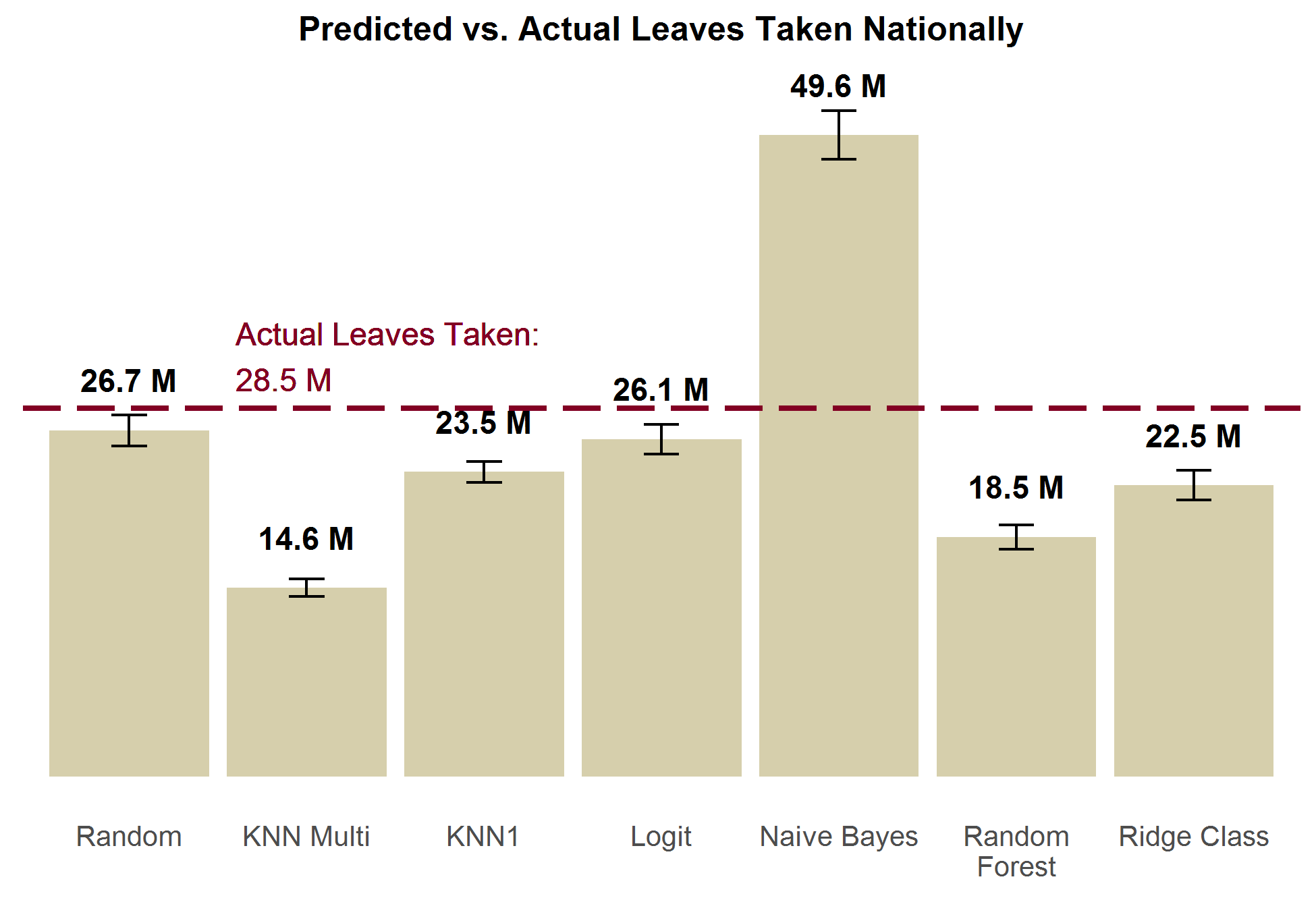
***Leave Takers.*** Exhibit X below shows the number of individuals nationally each method predicts to take at least one leave in a 12 month period, and compares that prediction to the actual number of 23.5 million leave takers. As we see, KNN1 hits closest to that mark; no other method captures the actual value within their 95% confidence interval. Logit, KNN multi, random forest, and ridge class all substantially understate the number of leave takers, while Naïve Bayes overstates it (with a prediction of 26.2 million leave takers).

**Exhibit 6.**



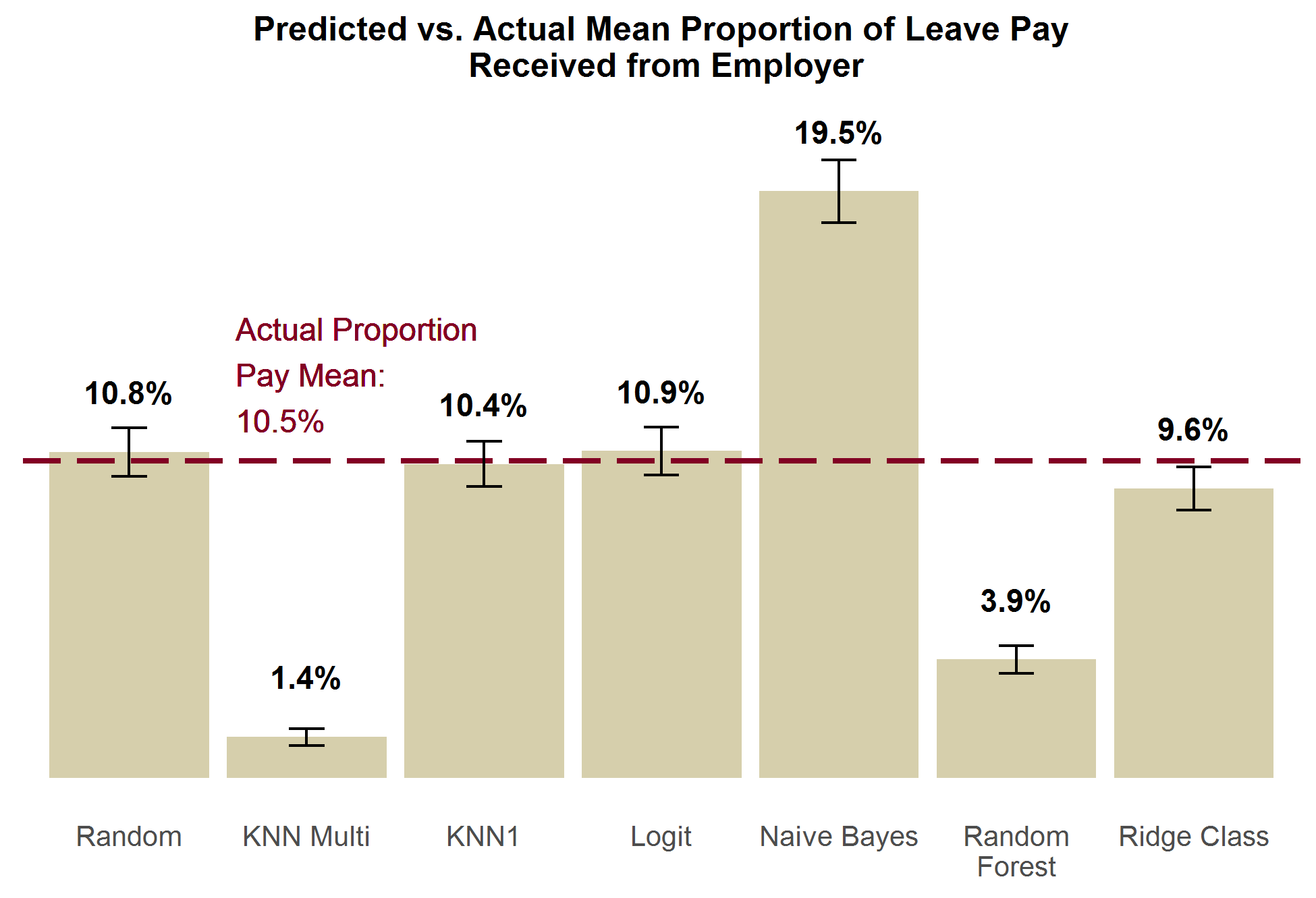
***Leaves Taken.*** Exhibit X below is slightly different from Exhibit X above. It measures the total predicted *leaves* rather than *leave takers*. There are a number individuals who require multiple leaves, and so these numbers are not the same. Correspondingly, there are more actual leaves taken (28.5 million) than actual leave takers (23.5 million). While no method captures the true value within their estimate’s confidence interval, logit comes closest with a small underestimation of 26.1 million leaves. KNN1 is next with 23.5 million, and ridge class is third with 22.5 million. KNN multi and random forest more drastically understate leaves taken. Naïve Bayes drastically overpredicts number of leaves; predic ting 49.6 million leaves.

**Exhibit 7.**



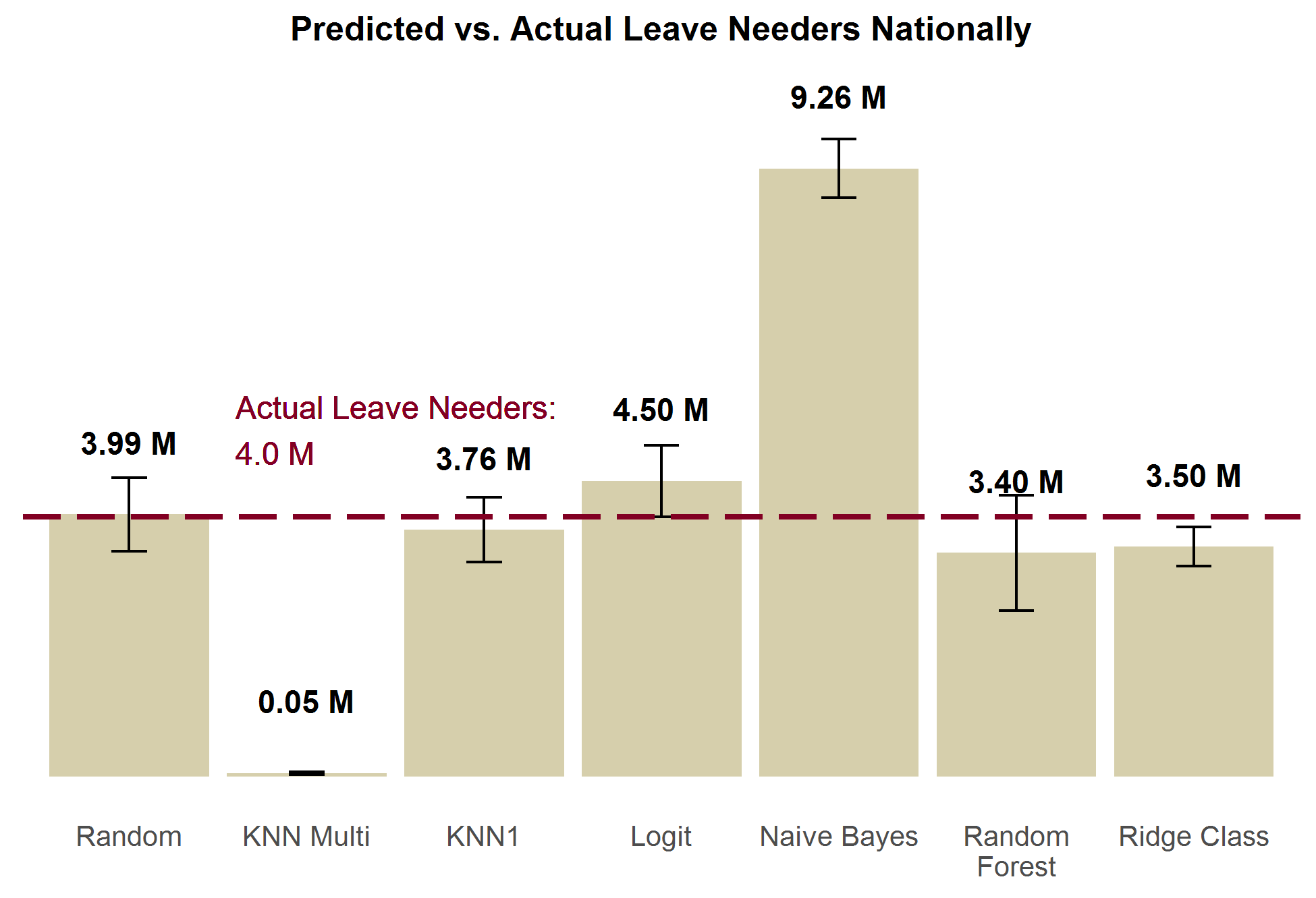
***Proportion of Pay Received from Employer.*** For the proportion of pay received, we compare the average mean predicted for each method in Exhibit X below. On average, individuals actually receive about 11% of their wages from their employers. KNN1 and logit methods both have estimates close to the actual value; both capture the actual mean within their estimate’s confidence interval. Ridge class is close with a mean prediction of 9.6%, but undershoots the mean value by a small amount. Random forest and KNN1 both drastically underestimate the proportion of pay received, while Naïve Bayes drastically overstates the proportion of pay received.

**Exhibit 8.**



***Leave Needers.*** Exhibit X below displays the predicted versus actual leave needers. There were actually 4 million leave needers in 2011 according to the FMLA survey. KNN1, logit, ridge class, and random forest methods all come close to properly estimating this. KNN multi drastically understates leave needing, and estimates just a handful of individuals will need leave. Naïve Bayes drastically overstates the number of leave needers.

**Exhibit 9.**



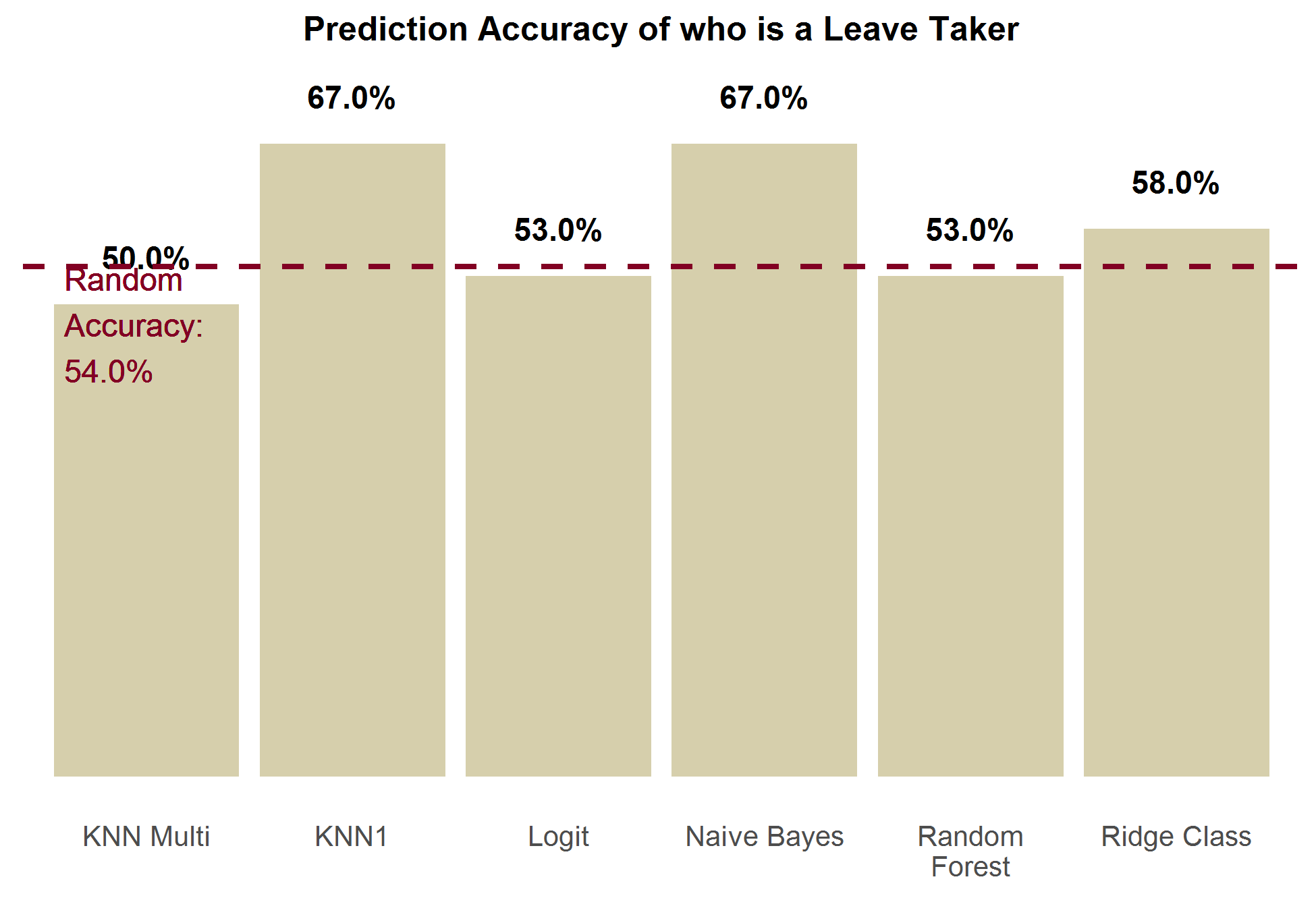
### X.2.3 FMLA-to-FMLA Individual-Level Performance

The previous subsection looking at aggregate performance of each method; how accurately they predicted “*how many* individuals took leave/needed leave/etc.” But we also want to know how well these models predict “*who* took leave/needed leave/etc.” To find this out, we also tested how accurate the predictions of the methods were at the individual-level, as well as their precision and recall. In this section, we also compare methods’ performance against random draws as baseline performance. The improvement from random draws is illustrative of the marginal gain we have achieved by using the given imputation method.

In this section, the rank order of the methods by different measures is significantly more heterogeneous than the results from the previous two section’s tests. The lack of consistency in these results leaves is a contrast of the conclusive evidence of KNN1’s superiority from the previous two sections. What is most consistent and instructive from these tests is the relatively modest gains in performance models typically exhibit over random draws.

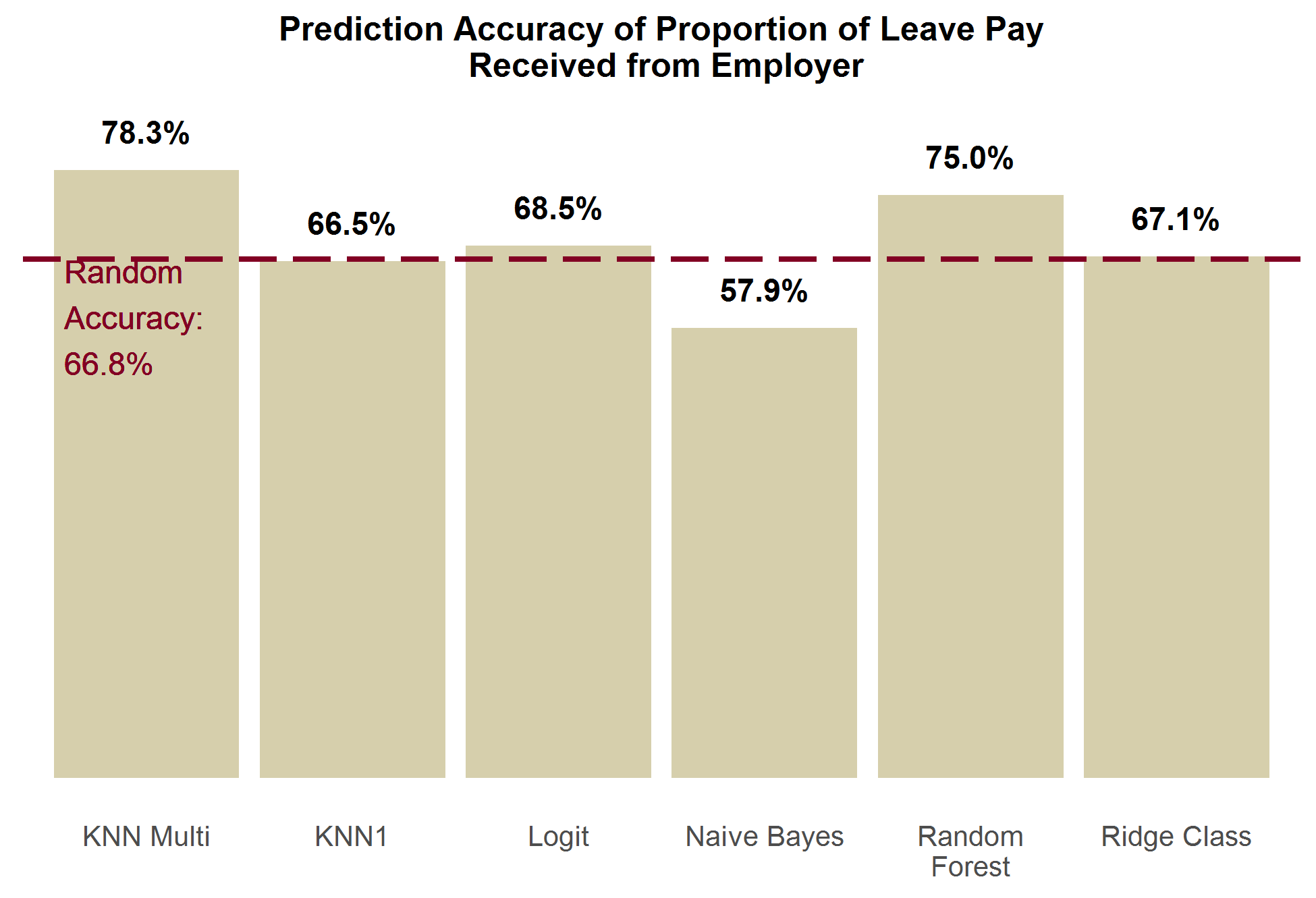
***Leave Takers.*** Exhibit X looks at the overall accuracy of each model at predicting leave takers, compared to random draws. Only KNN1 and Naïve Bayes are appreciably better than random draws. Even these two methods still only show modest improvement over random draws, and still often make errors. This is a strong indicator that conditional independence does not hold; there are unobservable characteristics related to leave taking, which leads to biased predictive models. We are only able to use the limited set[[1]](#footnote-1) of overlapping demographic characteristics between the FMLA and ACS surveys; which belies the true complexity of leave taking decisions. These results drive our recommendation to use our model primarily to answer population-level (“how many take leave?”) research questions, and to view answers to individual-level (“who takes leave?”) research questions with caution.

**Exhibit 10.**



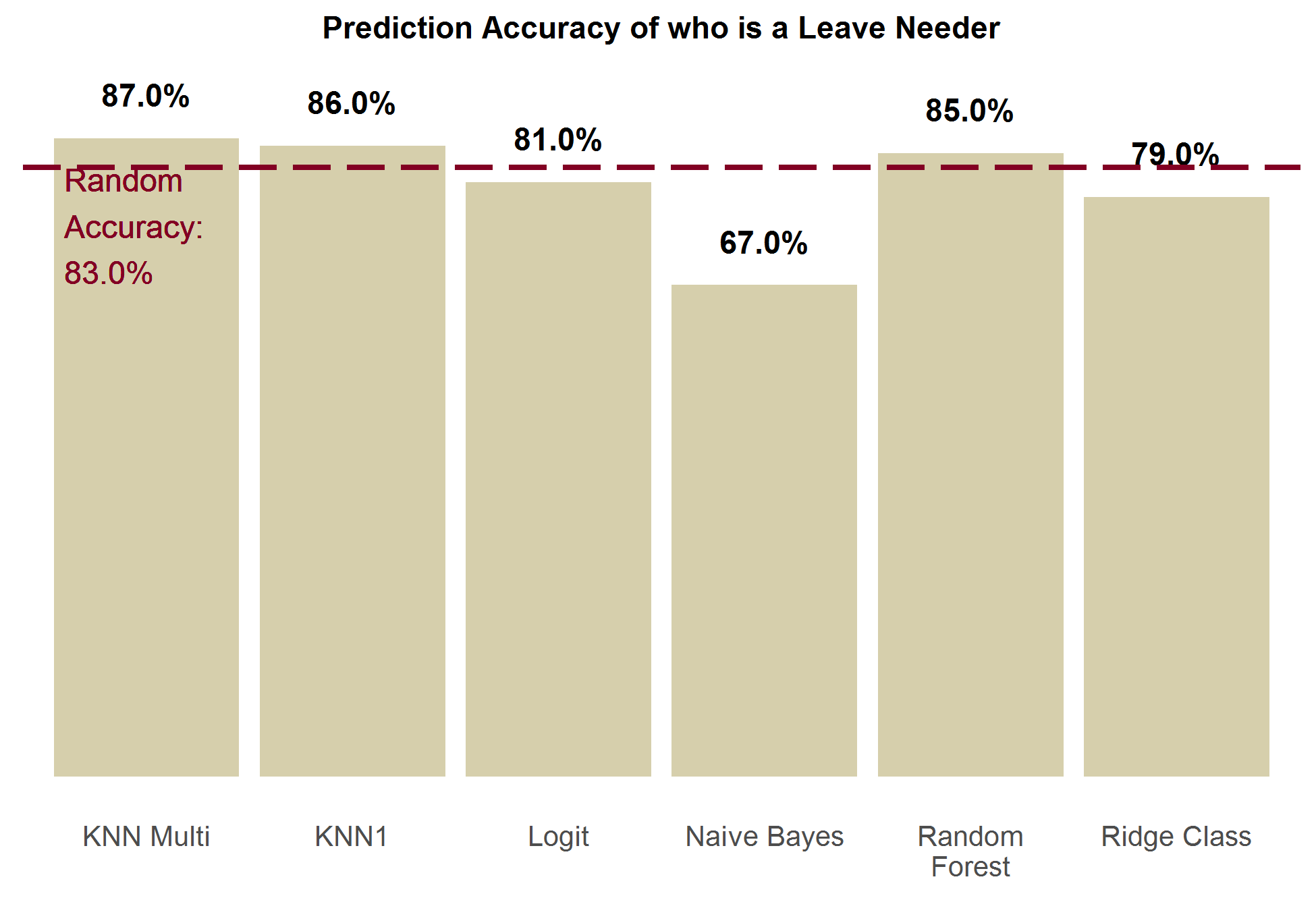
***Proportion of Pay Received from Employer.*** Exhibit X shows how each imputation method performs at predicting the correct proportion of leave pay received from their employer, compared to each other and to random draws. This is the proportion of individuals a method predicted the correct proportion of pay value for (out of the 7 possible values). KNN multi is the most accurate, closely followed by random forest. The other 4 methods are all comparable or worse than random draws. Naïve Bayes was the only method to perform worse than random draws.

**Exhibit 11.**

******

***Leave Needers.*** Exhibit X shows that no method does much better than random accuracy, and half of the methods actually do worse than random. KNN multi is the best performing method, but only gets an additional 4% better accuracy than random guessing.

**Exhibit 12.**



# Appendix X. Parameters Used to Simulate

# Actual Paid Leave Programs

The table below presents the list of parameters that were altered from state to state to conduct simulations of their actual state paid leave programs. These are based on the real-world eligibility requirements for these state’s programs

| **Parameter** | **Description** | **California** | **New Jersey** | **Rhode Island** |
| --- | --- | --- | --- | --- |
| bene level | Proportion of pay received as part of program participation (also known as the wage replacement rate) | 0.55 | 0.66 | 0.6 |
| maxlen own | Max number of days **own health** benefits can be claimed within a 12 month period | 260 | 130 | 150 |
| maxlen illspouse | Max number of days **ill spouse** benefits can be claimed within a 12 month period | 30 | 30 | 20 |
| maxlen illchild | Max number of days ill **child** benefits can be claimed within a 12 month period | 30 | 30 | 20 |
| maxlen illparent | Max number of days ill **parent benefits** can be claimed within a 12 month period | 30 | 30 | 20 |
| maxlen matdis | Max number of days **maternal disability** benefits can be claimed within a 12 month period | 260 | 130 | 150 |
| maxlen bond | Max number of days **child bonding** benefits can be claimed within a 12 month period | 30 | 30 | 20 |
| maxlen DI | Max number of days **maternal disability** **and own health benefits combined** can be claimed within a 12 month period | 260 | 130 | 150 |
| maxlen PFL | Max number of days **child bonding and ill child/parent/spouse benefits combined** can be claimed within a 12 month period | 30 | 30 | 20 |
| maxlen total | Max number of days **benefits of all types combined** can be claimed within a 12 month period | 260 | 130 | 150 |
| week bene cap | Max weekly benefits that can be collected as an absolute value | 1216 | 594 | None |
| week bene cap prop | Max weekly benefits that can be collected as a proportion of the state’s mean weekly wage | None | None | 0.85 |
| earnings | Eligibility Requirement - Minimum earnings (in dollars) within past 12 months | 300 | 8400 | 11520 |
| Own elig adj | Program eligibility adjustment factor for **own health** leave | 1 | .7 | 1 |
| Matdis elig adj | Program eligibility adjustment factor for **maternity disability** leave | 1 | .7 | 1 |

1. Each model by default includes all of these overlapping variables. They are: gender, marital status, age, education, race, ethnicity, family income, and presence/absence of children. [↑](#footnote-ref-1)