ISYE 7406: Spring 2024

Team 100, Charles Kramer

Flipping Precincts: Voter Targeting with Public Data

Abstract: Can free, publicly available data be useful for voter targeting? This question is salient for small, less-funded campaigns for down-ballot offices—such as for state legislatures—that are nonetheless crucial in writing laws that affect everyday lives. I predict whether a precinct will flip parties based on its demographic characteristics, using freely available data from the Virginia Department of Elections and the US Bureau of the Census. I use Monte Carlo cross-validation on 5 machine learning models to choose the best model for the two types of flips (Republican->Democratic and Democratic->Republican). The winning models produce a substantial gain in accuracy over naive predictions, empowering down-ballot, grassroots candidates to campaign more effectively and improve their chances of winning elections—all at zero cost.

Code: https://github.com/Charlie-Kramer/precinct_flips

1. Introduction: voter targeting in low-funded campaigns

High-profile, well-funded political campaigns benefit from sophisticated analytics that target voters for efforts such as phone calls, candidate events, canvassing, mailers, or digital ads. These analytics are based on official voter registration data such as name, phone, address, and dates when the person voted. Commercial vendors merge this data with other data attempting to identify other key characteristics that predict voter propensities, characteristics like race, income, gender, and age, along with custom analytics to model the propensity to vote and the strength of party affiliation. Desilver (2018) found that these models are quite accurate in their predictions. Campaigns use these data to target campaign resources to the voters most likely to turn out and vote for their candidate.

Unfortunately, sophisticated analytics can be too expensive for small campaigns. Consider candidates for state legislature. In the 2023 campaigns for the Virginia state House of Delegates, median expenditures were just $149,000. For comparison, in the 2021-22 cycle, candidates for Congress spent about $3.3 million per race on average. But the less well-funded, state-level offices have a crucial impact on the everyday lives of citizens, legislating on issues such as gun safety, access to health care, and education.

I address this issue by finding pockets of voters that may be good targets for campaign efforts, using free data and tools. While voter-level data are not publicly available, data are available at the precinct level (for voting behavior) and census tract level (for demographic/economic characteristics). I exploit these data to find precincts that have a predilection for flipping—e.g. from majority-Republican to majority-Democratic votes or vice versa. Such precincts are likely good candidates for outreach efforts.

2. Data

To make the focus manageable, I examine the 2020 general election data and the swing from 2019 by precinct in the state of Virginia. I choose Virginia because it is a well known “purple” state without a strong overall partisan lean, but with substantial pockets of strength for both major parties.

What are the characteristics of likely swing voters? Pew (2021) identified the following characteristics of middle-of-the-road voters, their so-called Stressed Sideliners:

> A majority of Stressed Sideliners (56%) are women. Roughly six-in-ten (57%) are White, while 21% are Hispanic, 10% are Black and 5% are Asian. They generally look similar to U.S. adults overall in terms of age: 18% are under the age of 30, 34% are between 30 and 49, 31% are 50 to 64, and 17% are 65 and older.…About one-in-four (43%) live in lower-income households, with just 10% living in upper-income households.

That is, gender, race, age, and income are key characteristics (see also Pew’s analysis of independent voters (Pew 2019)). Other characteristics that are can segment party preferences include whether or not one is foreign-born, whether one has broadband internet access (a proxy for rural/urban), and whether one is eligible to access the public safety net.

Accordingly, I drew the following variables from the U.S. Census American Community Survey (ACS) for 2018, using 5-year estimates (the 3-year and 1-year estimates are more timely, but less comprehensive, and do not include figures for areas with population as small as a census tract). The one-year lag (2018 vs 2019) accounts for lag in ACS data.

- The percentage of male residents
- The percentage of white residents
- The median age
- The percentage of foreign-born residents
- The percentage of residents living under the poverty line
- The percentage with broadband access
- The percentage that qualify for Medicaid (U.S. medical care assistance for the needy)

The census data are then matched to the closest voting precinct, using geographic shapefiles for census tracts and voting precincts. I use distance between centroids to measure proximity, and accordingly convert from a geographic geometry (latitude/longitude based) to a non-geographic geometry to calculate distances (distances based on degrees can be highly inaccurate). Each voting district is then assigned one of four classifications based on vote share totaled across all offices on the ballot: (1) majority Democratic in both 2019 and 2020 (DD), (2) majority Democratic in 2019 but majority Republican in 2020 (DR), (3) majority Republican in 2019 but majority Democratic in 2020 (RD), and (4) majority Republican in both 2019 and 2020 (RR). This classification is of most practical use as the vote share in the previous year is known prior to the election in question, and a campaign analyst will be interested in targeting precincts with a propensity to swing from the opposition party to their own candidate's party. That is, a Republican candidate would be interested in precincts that were majority-Democratic votes in the past, but could be converted to majority-Republican (discussed in more detail below).

One obvious shortcoming of the analysis is the lack of a campaign effort variable encompassing canvassing, phone banking, advertising, and other outreach. That is, some voting flips are undoubtedly caused by a persuasive and comprehensive campaign. Accordingly, effort would be an important control variable (indeed, the premise of the exercise is that effort could profitably be directed at particular precincts). Unfortunately, relevant data on this account do not exist. The Federal Elections Commission requires disclosures on expenditures both by the campaign and by unaffiliated entities (independent expenditures). However, these are not broken down by precinct (indeed, this would seem impossible for spending on say social media ads), and they do not include data on phone banking, canvassing, or public appearances by the candidate.

All variables except the target classification are standardized before processing. The dataset had no evident outliers (see Exploratory Data Analysis). I dropped precincts for which data were not available for either 2019 or 2020; it did not seem sensible to try to impute missing vote tallies given the volatility of the data. For the ACS data, some tracts had missing values for age due to an insufficient number of observations; for these, I replaced age with the median for all other tracts. This process yielded data on voting and demographics for 2,430 precincts.

3. **Exploratory Data Analysis**

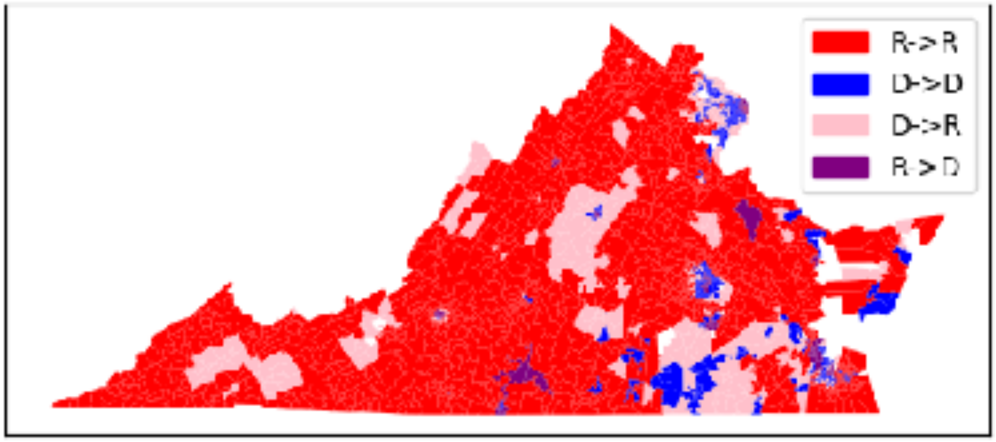
An examination of the data on precincts flips shows that the categories in the voting data are highly unbalanced (Table 1). For instance, only 8.9 percent of districts flipped from Republican-majority to Democratic. The largest categories are precincts that did not change parties; 52.6 for Republicans and 20 percent for Democrats.

Table 1: Frequency of Precinct Flips, Virginia 2019-2020			
		Unconditional P(2020, 2019)	
		2020	
		D	**R**
2019	**D**	20.0	18.5
	R	8.9	52.6
		Conditional P(2020 | 2019)	
		2020	
		D	**R**
2019	**D**	52.0	48.0
	R	14.5	85.5

This structure complicates predicting flips given knowledge of the base-year outcome (second panel in Table 1). Conditional on knowing the outcome in the previous year, one could predict a flip from Democratic to Republican at random with only 50 percent certainty, or a flip from Republican to Democratic with 14.5 percent probability.

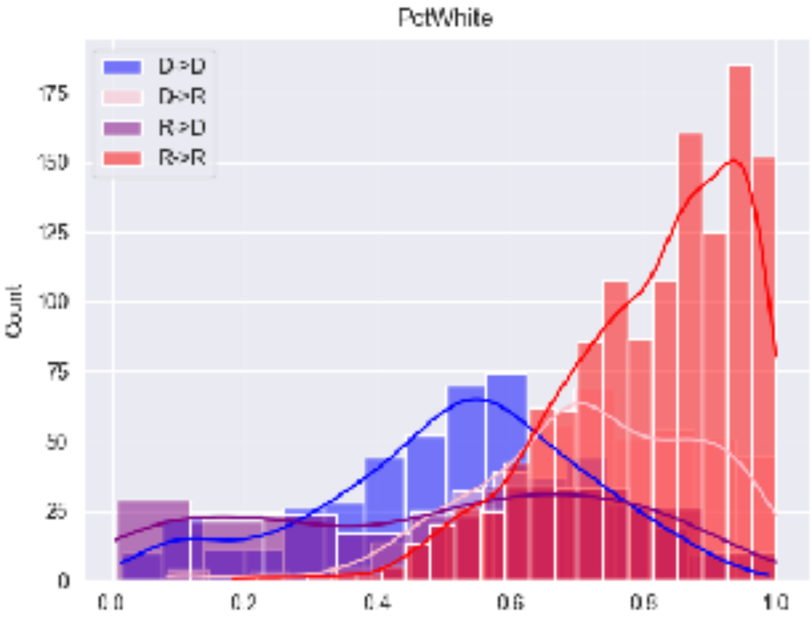
However, flipping districts tend to cluster (Figure 1)—see for example the pink clusters (D->R) in Southwest and Central Virginia, and the purple clusters (R->D) in Northern Virginia and Southern Virginia. This is desirable because campaigns can take advantage of geographic proximity for candidate appearances and canvassing. The tendency to cluster is confirmed by join count statistics (Rey et al, 2023), which yield a p value of 0.001 for the null hypothesis that flips are randomly distributed (Annex Table 1).

Figure 1. Map of Precinct Flips, Virginia 2019-2020



Another question is whether our seven explanatory variables differ across categories, so that they can distinguish outcomes. Figure 2 shows the distribution of Percent White across precincts, segregated by flip category (RR, RD, DR, DD). The distributions are markedly different. In particular, precincts with a higher share of white persons are more likely to remain Republican (RR) or flip from Democratic to Republican (DR). K-sample Anderson-Darling tests (Scholz and Stephens, 1987) for equality of the four distributions rejected the null at the 0.001 percent significance level (see Annex Table 1, and Annex Figure 1 for the distributions and tests for other demographic variables).

Figure 2. Distribution of Percent White by Flip Category



4. Modeling

This is two problems, not one

As noted, in the data, there are four possible labels: RR, RD, DR, and DD, corresponding to the predominant party in 2019 and 2020 in that precinct. In practice, however, only half of these labels apply to a particular precinct. For instance, if the precinct is R in 2019, the possible outcomes are only RR and RD. The outcomes DR and DD cannot occur because the precinct is not D in the base year.

Thus, one would not want to model all four outcomes for every precinct—two of the four outcomes are impossible. Moreover, an analyst working for a Republican candidate would be interested in precincts that can be flipped from majority Democratic to majority Republican, so they would be interested in predictions for precincts that are currently majority Democratic.

Accordingly, we have two problems rather than one. Accordingly, I split the data into observations for which the precinct is majority Republican in 2019 (*Base-R*; outcomes RD and RR) and those that are majority Democratic (*Base-D*; outcomes DR and DD).

Addressing unbalanced categories

After splitting by base, the issue of unbalanced categories remains. To address this, I use Synthetic Minority Oversampling Technique (SMOTE; Ching 2021) to generate more balanced training data—this avoids the pitfall of the machine learning techniques collapsing to a naive prediction of the majority class. That is, consider a two-class dataset that has 90 percent class A and 10 percent class B; a naive model that predicted only class A would be nominally 90 percent accurate but not useful for predicting class B. I chose K=8 neighbors for SMOTE on the basis of 100 Monte Carlo cross-validation runs, with the objective to maximize the predictive power (balanced accuracy) of an ADABoost model on the SMOTEd dataset.

The unbalanced nature of the outcomes also influences the choice of objective for evaluating candidate models and model parameterizations. I use balanced accuracy:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Sensitivity is the true positive rate and Specificity is the true negative rate. I use this as the overall objective in choosing the ‘best’ model in fitting, as it accounts for imbalanced classes. This is useful as using only sensitivity (say) can cause the model to focus on predicting the majority class, at the expense of the minority class. Balanced accuracy accounts for unbalanced data by measuring specificity as well as sensitivity.

Upon finding the ‘best’ model I also report in my findings Matthews Correlation and F1 score as well (both used for assessing predictive power in the presence of unbalanced data):

$$Matthews\ Correlation = \frac{TN\,TP - FN\,FP}{\sqrt{(}(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$$

$$F1\ Score = \frac{2PrecisionRecall}{Precision + Recall}$$

Where TN= true negative, TP=true positive, FN=false negative, FP=false positive, Precision=TP/(TP+FP), and Recall = TP/(TP+FN).

Assessing usefulness of the results

Based on my experience working on a state legislature campaign, the objective is to improve on the decisions that would be made in the absence of model estimates of flip probabilities. Consider the following example. Suppose that we are working for a Democratic campaign and thus are interested in predicting which Base-R precincts may flip. That is, given a set of precincts that are currently R, we want to know which of them are likely RD.

Given the very limited funding that is the hallmark of down-ballot campaigns, the appropriate benchmark for usefulness is “better than nothing”—because “nothing” (beyond very basic data) is what the campaign may very well have at hand. This benchmark is merely the conditional probability observed in the data—e.g. the probability of predicting a flip correctly in the absence of any information other than the base-year outcome. From Table 1, this is 14.5 percent for Base-R. A model estimate that improves substantially on this is useful. In particular, I calculate (for instance, for Base-R precincts)

$$Gain(m, R) = P(DR|Base = R, model = m) - P(DR|Base = R)$$

And then test the null hypothesis:

$$H_0 : Gain(m, R) = 0$$
$$H_1 : Gain(m, R) > 0$$

Or equivalently,

$$H_0 : P(DR|Base = R, model = m) \leq P(DR|R)$$
$$H_1 : P(DR|Base = R, model = m) > P(DR|R)$$

5. Model fitting and results

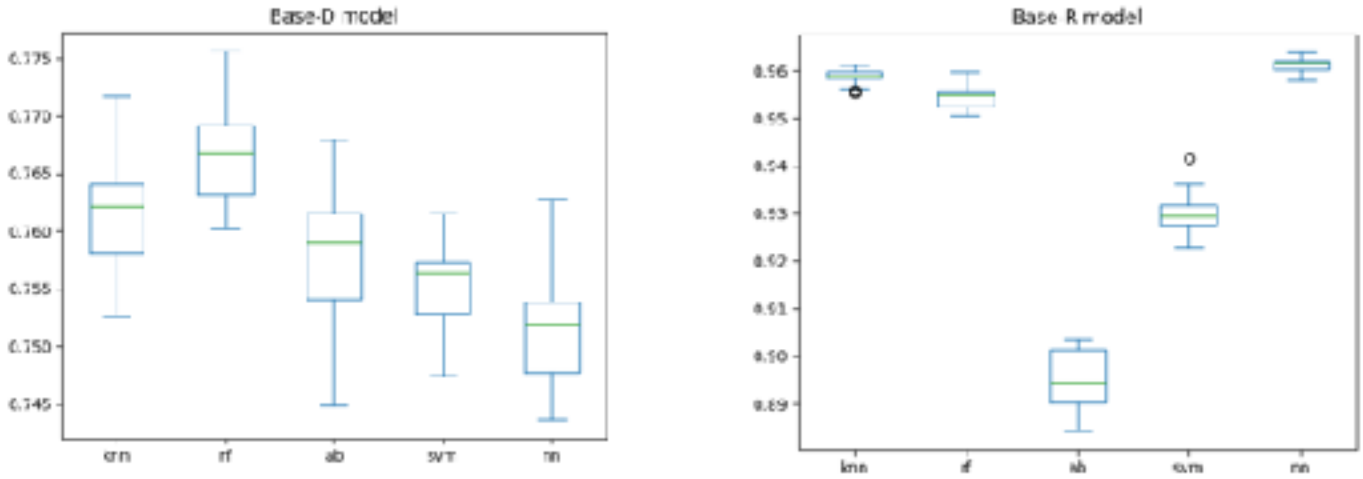
Modeling followed the process shown below, with 30 Monte Carlo SMOTE repetitions. Within each of those repetitions, 10-fold cross validation of each of the five models selects the best parameters, then re-fits the model with those parameters and evaluates its performance. In a last step, the best of the 30 parameterizations is used for a final evaluation. Given the modest size of the data set and the unbalanced data, I did not hold out a single test set for the exercise through the last step.

```
For base in {base-D, base-R}:
    Do 30 times:
            Randomly split sample into train/test (80/20)
            Use SMOTE to generate balanced categories for training data
            For model in {K Nearest Neighbors, Random Forest, ADABoost, SVM, Neural Net}:
                    Choose parameters based on 10-fold grid CV (Table 1) on training data
                    Re-fit model with chosen parameters, generate balanced accuracy on test
                    If balanced accuracy > previous results:
                            Save results as best model parameters (see Annex Table 2)

    For model in {K Nearest Neighbors, Random Forest, ADABoost, SVM, Neural Net}:
            Retrieve parameters obtained from best results of above loop
            Randomly split train/test
            SMOTE training set
            Re-fit model to train, generate diagnostics on test (see Table 2)
```

Figure 3 shows the results of the MC CV loops (Annex Table 2 shows the CV grid and parameters associated with best results for each model). Random Forest (rf) performs best overall on base-D problems, with a higher central tendency and a distribution that has most of its mass above those for other methods. That said, the differences among them are not large (second or third decimal place). The base-R dataset shows that Neural Net (nn) similarly outperforms, although the KNN and RF models produce very good results (balanced accuracy above 90 percent). Formal t-tests confirm that the mean results for Random Forest (base-D)

Figure 3. MC CV Results—Balanced Accuracy



and Neural net(base-R) are better than those for the competing models (Annex Table 1).

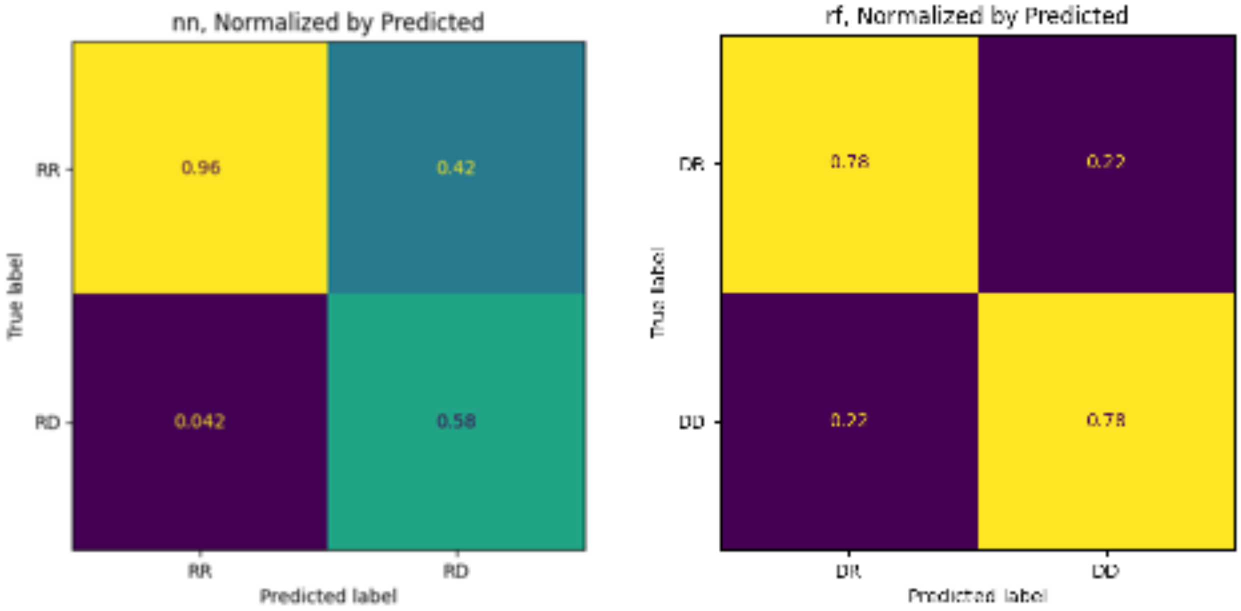
Table 2 shows the final balanced accuracy, F1 score, and Matthews correlation for the best parameterizations re-fit to a new test/train split. Table 2 shows that the Random Forest model is superior to the others for both base-R and base-D models, with F1 and balanced accuracy in the 70 to 80 percent range. However, the overall differences across models are modest, and no model dominates all others across all performance metrics.

Table 2: Scores from Re-fitting and Re-Testing Best Model Parameterization

Model	**F1 score**	**Matthews**	**Balanced Accuracy**
	Base-D model		
KNN	0.7807	*0.5619*	*0.7809*
Random Forest*	*0.7853*	0.5613	0.7806
ADABoost	0.7539	0.4971	0.7485
SVM	0.7789	0.5507	0.7753
NN	0.7254	0.4329	0.7162
	Base-R model		
KNN	0.7059	0.6557	*0.8659*
Random Forest	*0.7253*	*0.6765*	0.8475
ADABoost	0.6666	0.6145	0.8616
SVM	0.6857	0.6331	0.8601
NN*	0.6602	0.6006	0.8373
	Best results in each category are italicized. * = best model based on CV t-test.		

Figure 4 shows the confusion matrices from final re-fitting of the models, normalized by the predicted category. The probability that a precinct labeled RD will actually flip is 58 percent, while the probability that a precinct labeled DR will flip is 78 percent. This compares to benchmark probabilities (Table 1) of 14.5 percent and 48 percent respectively. Accordingly, the gain is 43.5 percentage points for Base-R precincts and 30 percentage points for Base-D precincts. The test statistics for the respective null hypothesis that gain is zero are 47.7 and 18.3 respectively, both of which reject the null at usual significance levels.

Figure 4: Confusion Matrices, Best Models



5. Conclusions

The results show a substantial increase in accuracy of prediction compared to a naive prediction. That is, there is a 30 percentage point improvement in targeting for currently Democratic precincts, and a 43.5 percentage point increase in accuracy for currently Republican precincts. This would imply substantial resource savings for campaigns in targeting efforts such as candidate appearances, mailers, and canvassing.

Consider the following simple example. Suppose a Republican campaign wants to target Democratic districts to flip to Republican. It can spread its campaign resources across all currently-Democratic precincts, or it can focus on those that the model predicts will flip. Denote resources allocated to precinct *i* by X_i and the increase in votes per allocation, conditional on flip category, as $V(X_i|C)\ where\ C \in \{DR, DD\}$. The expected increase in votes for a given allocation is then $V(X_i|DR)P(DR) + V(X_i|DD)P(DD)$. Suppose that solidly Democratic precincts do not respond to allocation of resources by Republican campaigns: $V(X_i|DD) = 0$. Then the expected increase in votes is merely $V(X_i|DR)P(DR)$ and the increase in expected votes gained for that precinct from modeling is $V(X_i|DR)[P(DR|model) - P(DR|no\ model)]$ or in this case, 30 percent. Given the often narrow margins by which campaigns are won or lost, this is a sizable increase—and (as noted above) statistically significant.

Accordingly, this type of modeling can give campaigns a considerable improvement in effectiveness with no attendant increase in expenditure, using public data and open-source tools. This levels the playing field with larger, better-funded campaigns. Ultimately it improves the ability of grassroots campaigners that lack major funding to connect effectively with voters and win elections.

7. References

Ching, 2021, Introduction to Synthetic Minority Over-sampling Technique and its Implementation from Scratch, Towards Data Science (Medium).
Desilver, Drew (2018), Q&A: The growing use of 'voter files' in studying the U.S. electorate, Pew Research Center.
Pew Research Center (2019), Political Independents: Who They Are, What They Think.
Pew Research Center (2021), Beyond Red vs. Blue: The Political Typology.
Rey, Sergio J., Dani Arribas-Bel and Levi J. Wolf (2003), "Global Spatial Autocorrelation," in Geographic Data Science with Python.
F.W. Scholz, and M.A. Stephens (1987), "K-Sample Anderson-Darling Tests," Journal of the American Statistical Association, Vol. 82. No. 399, pp. 918-924.

Annex Table 1. Miscellaneous Statistical Tests

Join-count tests

	"BB"	"BW"
base-D	0.001	1.0
base-R	0.001	1.0
	"BB" tests the null that the number of similarly- labeled neighbors is not statistically different from random assignment ("BW" for differently labeled).	

CV: equality of means tests

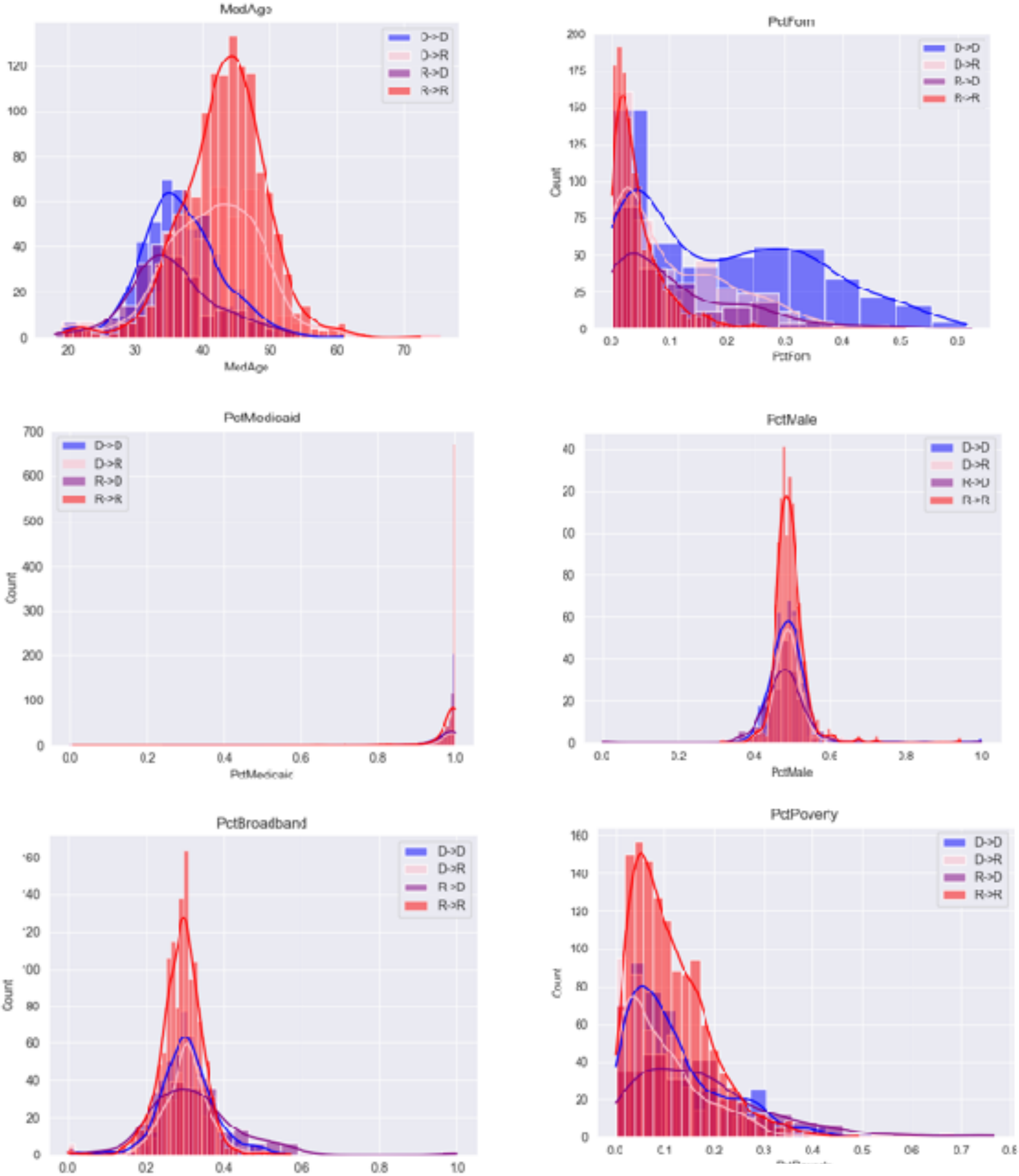
Base D: RF-KNN	4.112	0.0001
Base R: NN-RF	13.113	0.0000
	Tests that the model's average score is statistically larger than the average score for the next-best model across 30 MCCC runs; does not assume same variance for both populations	

K-sample Anderson-Darling tests		
Variable	**Statistic**	**P-value**
% Male	14.9	0.001
% White	336.0	0.001
Median Age	190.0	0.001
% Foreign Born	199.8	0.001
% Poverty	38.7	0.001
%Broadband	26.8	0.001
% Medicaid	24.1	0.001

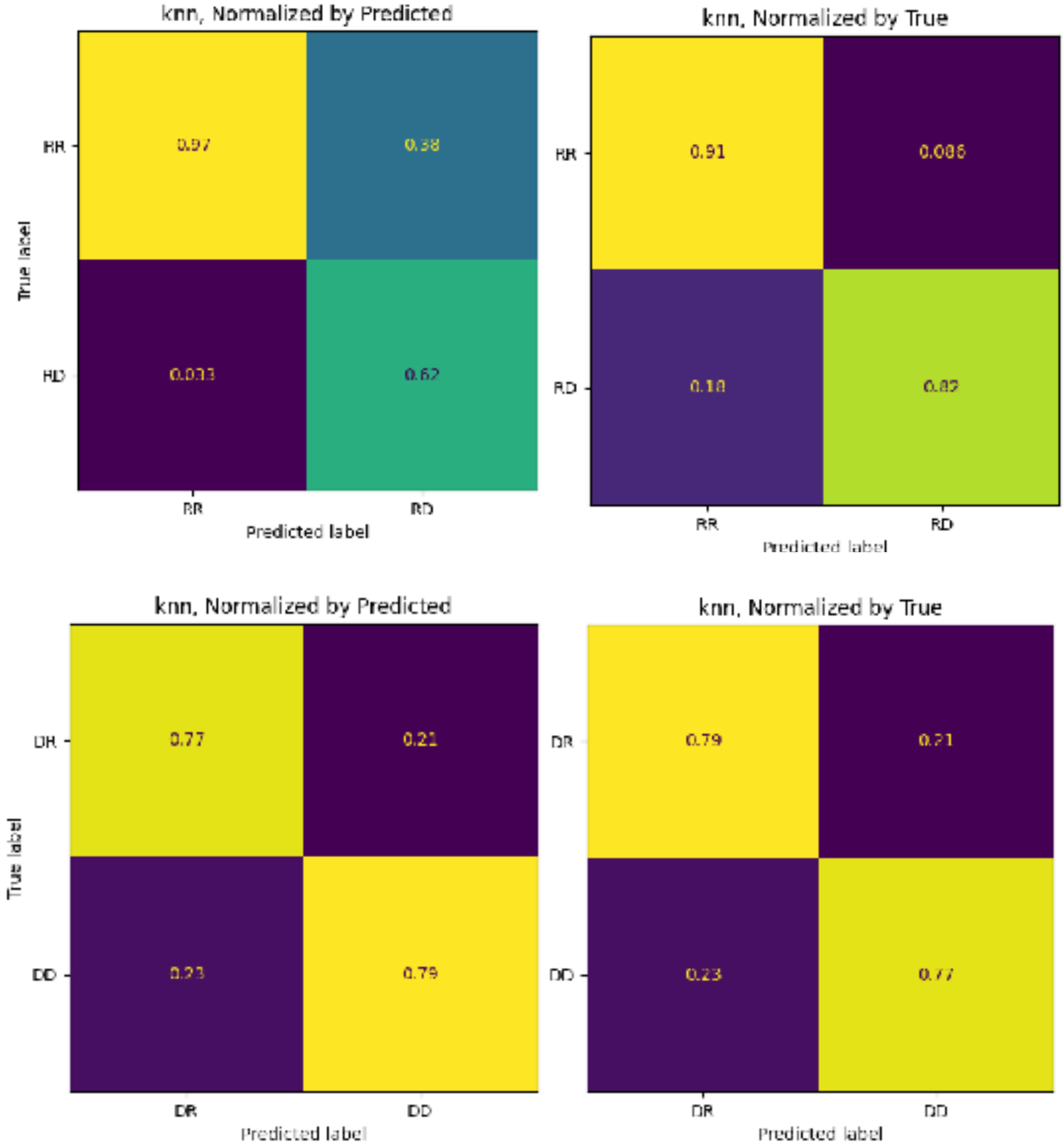
Annex Table 2. Model CV Parameterization

Model	**Key Parameters (CV grid; bold = parameter chosen by 30 MC x 10-fold CV)**
Submodel	Base-D model
KNN	Number of neighbors (1-20; **17**); P-exponent on distance metric (**1**, 2); weights (**uniform**, distance)
Random Forest	Number of estimators (20, 30,,..,200; **120**), criterion (gini, **entropy**, log_loss), minimum samples for split (**2**,4,6), minimum samples per leaf (1,3,**5**)
ADABoost	Number of estimators (5, 10, 15,..,100; **55**), learning rate (.25, .75, **1**, 2, 4),
SVM	C (**.5**,1,2), kernel (**linear**, poly, rbf, sigmoid), gamma (**scale**, auto)
NN	Activation(**tanh**, relu),hidden layer sizes(50,**100**,200), learning_rate(**constant**, adaptive)
Submodel	Base-R model
KNN	Number of neighbors (1-20, **4**); P-exponent on distance metric (1, **2**); weights (uniform, **distance**)
Random Forest	Number of estimators (20, 30,..,200; **140**), criterion (gini, **entropy**, log_loss), minimum samples for split (**2**,4,6), minimum samples per leaf (**1**,3,5)
ADABoost	Number of estimators (5, 10,..,100; **90**), learning rate (.25, .75, **1**, 2, 4).
SVM	C (.5,1,**2**), kernel (linear, poly, **rbf**, sigmoid), gamma (scale, **auto**)
NN	Activation(tanh, **relu**),hidden layer sizes(50, **100**, 200), learning_rate(**constant**, adaptive)

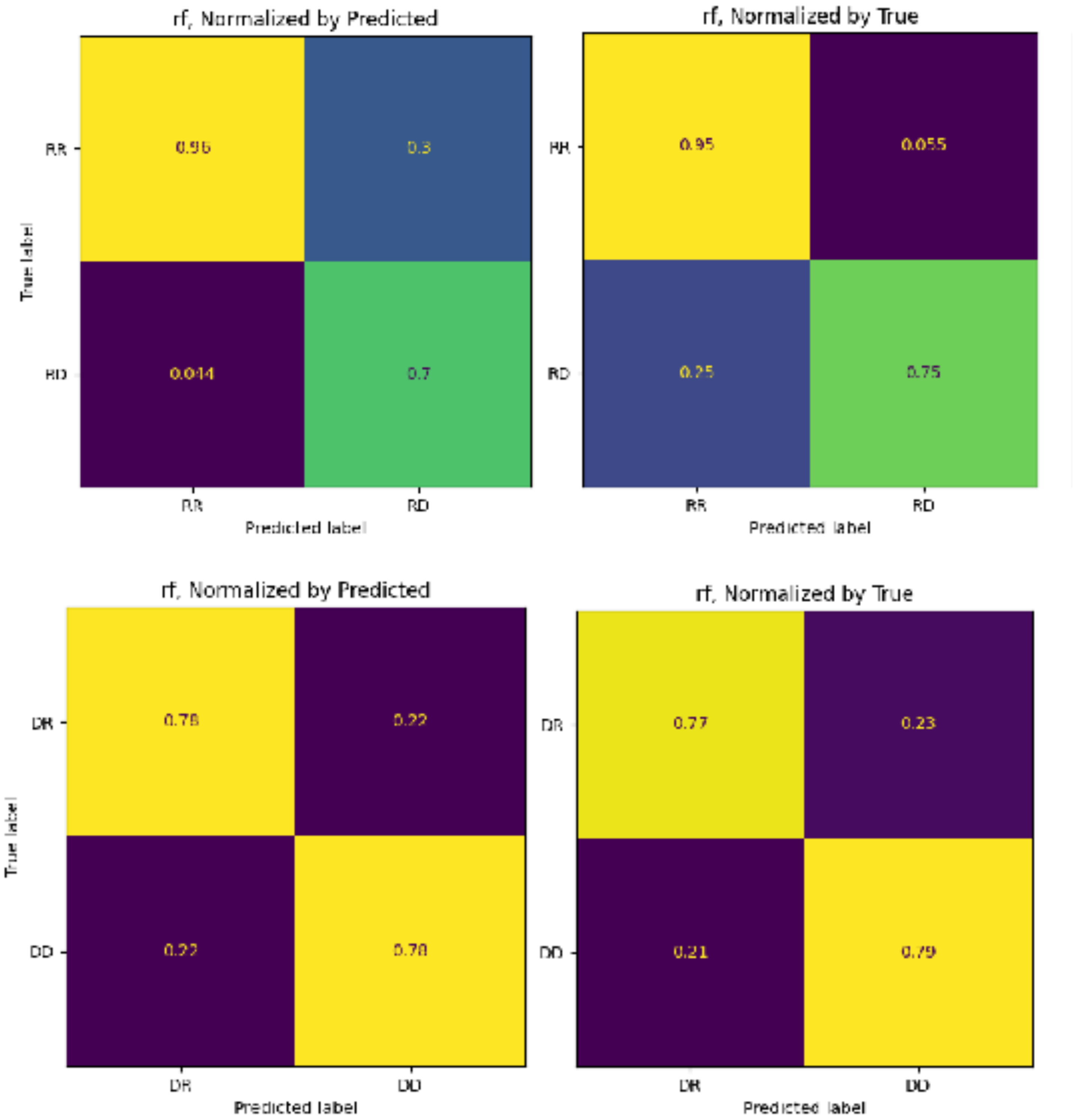
Annex Figure 1. Distribution of Demographics by Flip Category



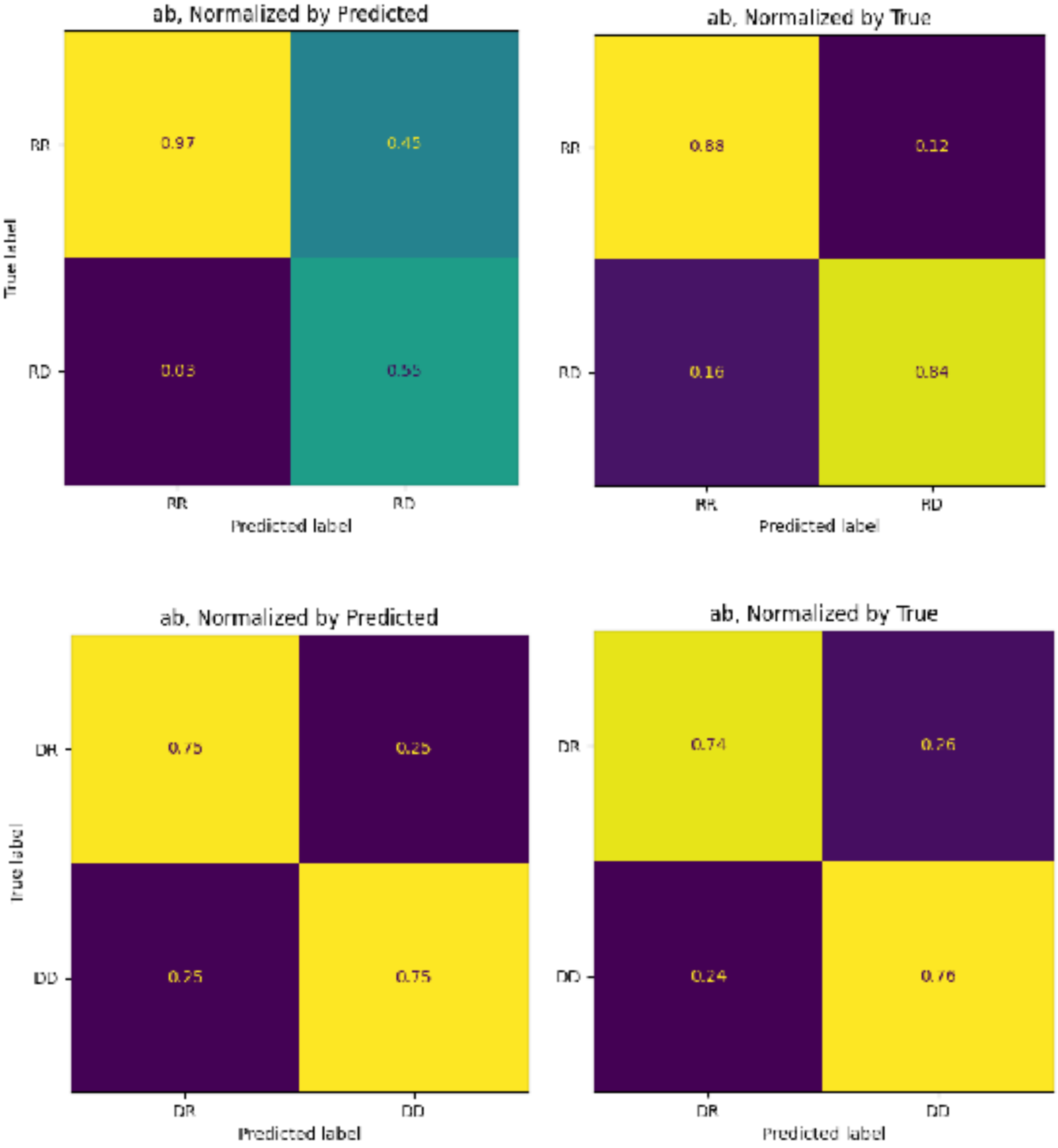
Annex Figure 2. Confusion Matrices, K-Nearest Neighbor Models, Base-R and Base-D



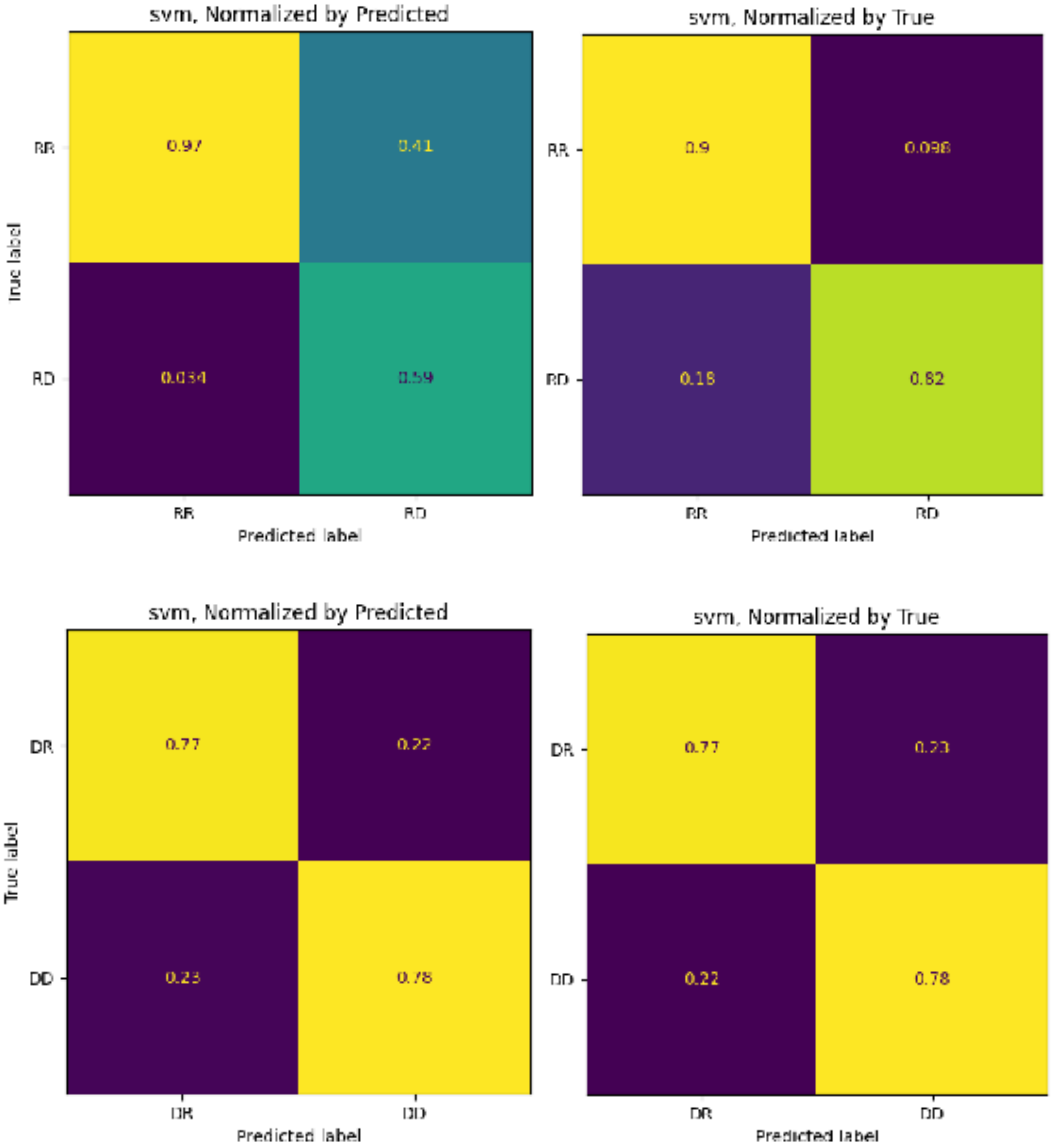
Annex Figure 3. Confusion Matrices, Random Forest Models, Base-R and Base-D



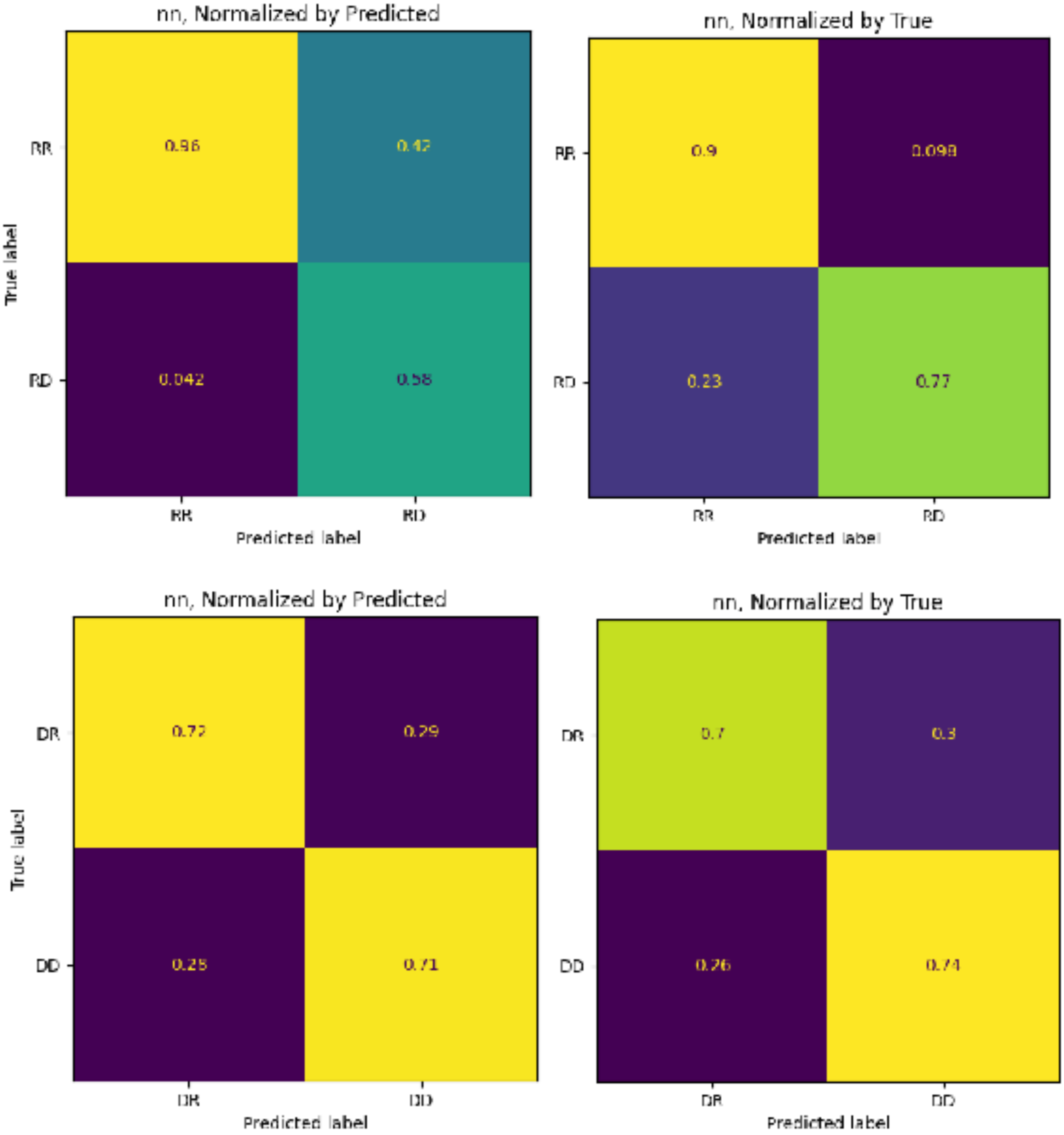
Annex Figure 4. Confusion Matrices, ADABoost Models, Base-R and Base-D



Annex Figure 5. Confusion Matrices, SVM Models, Base-R and Base-D



Annex Figure 6. Confusion Matrices, Neural Net Models, Base-R and Base-D


nn, Normalized by Predicted
True label
RR
RD
RR
RD
Predicted label
nn, Normalized by True
RR
RD
RR
RD
Predicted label
nn, Normalized by Predicted
True label
DR
DD
DR
DD
Predicted label
nn, Normalized by True
DR
DD
DR
DD
Predicted label