Classification Models: Predicting Moves with USAA Customer Data

Presentation

by Slava Bogoslovskiy

February 2015

Outline

- USAA Facts
- Challenge
- Data
- Models and results
- Feature selection
- Additional data

USAA Facts

- USAA was formed in 1922 by members of the Army Air Corps
- Their jobs were considered high risk, so they had to self insure
- All of the original “customers” were owners in an association, which is still true today
  - That’s why USAA calls its customers “members”
- To be a member you have to have honorably served in the US Military or be child or spouse of someone who is serving or has served.

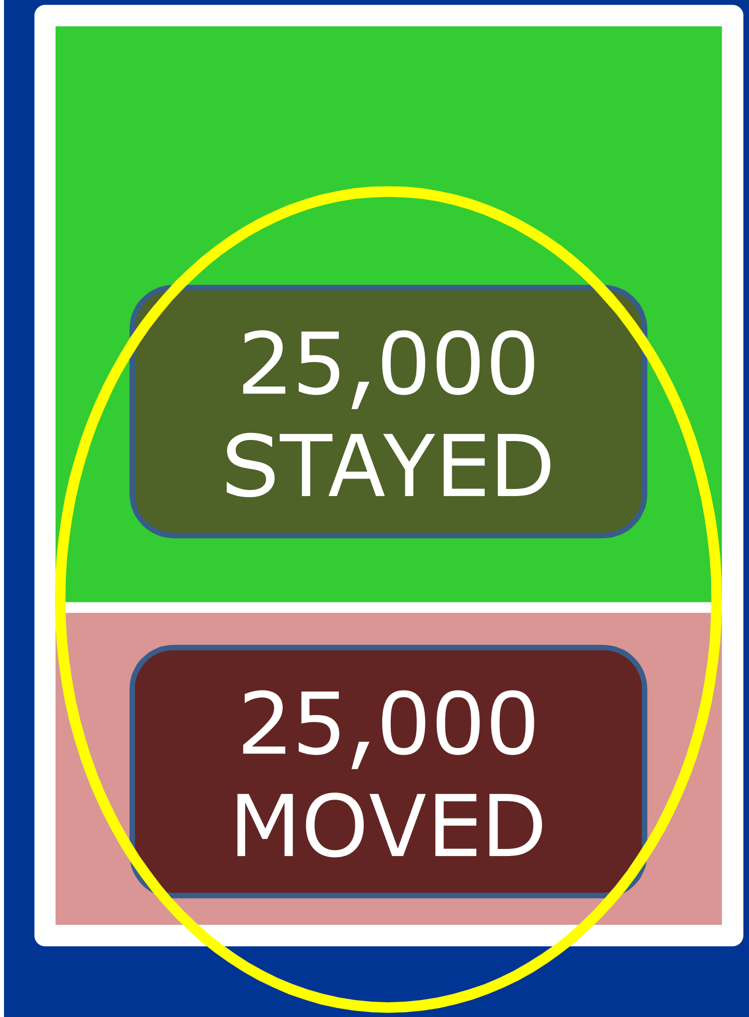
Challenge

USAA wants to predict when members are moving.

Why?

- People in the military move A LOT
- Each move is stressful, inconvenient, and often expensive
  - Predicting when members are moving can help take some of that stress off by automatically updating member information.
- Moving is associated with attrition
  - Predicting moves may help retain members

Data: Structure



Data

- Sample of 50,000 observations
- Each observation is an anonymized member with a vector of characteristics
- Data as of September 30, 2013
- 25,000 observations were randomly sampled from the set of all members who moved in October 2013
- The other 25,000 were randomly sampled from those who did not move
- Outcome (stayed or moved) is known

Data: Members' Characteristics

- 600 variables
  - Personal
    - age
    - education level
    - occupation
    - military status
    - number of kids
    - homeowner/renter
    - zip code, state
    - etc.
  - USAA-related
    - number of products
    - account balances
    - activity status
    - 3,6, and 12 month lags of some variables
    - etc.

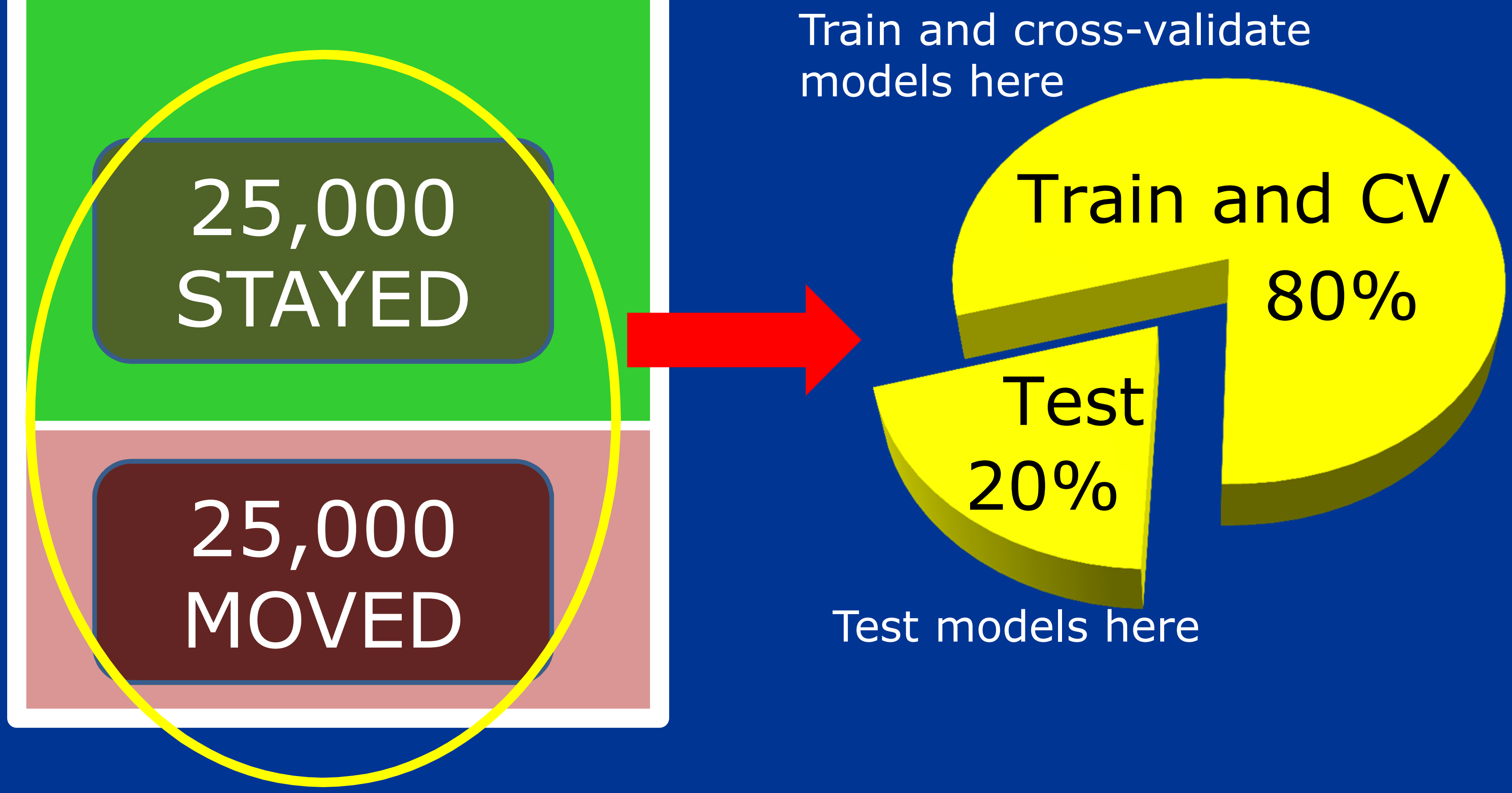
Data: Issues and Fixes

- Lots of missing data
  - Fix 1: impute missing data (NAs) using median or mode
  - Fix 2: delete variables with too many NAs
- Categorical variables with many levels
  - Fix: Reduce number of levels by grouping rare ones
- Poor variable description
  - Fix: data driven approach still feasible

Data After Cleaning

- 152 independent variables
- Some of them are highly correlated
  - That’s OK for predicting purposes
- Transform categorical variables into sets of dummy variables
- All numerical variables + all dummies for all categorical variables = set of 462 variables, call these “features”

Training and Testing Split



25,000 STAYED

25,000 MOVED

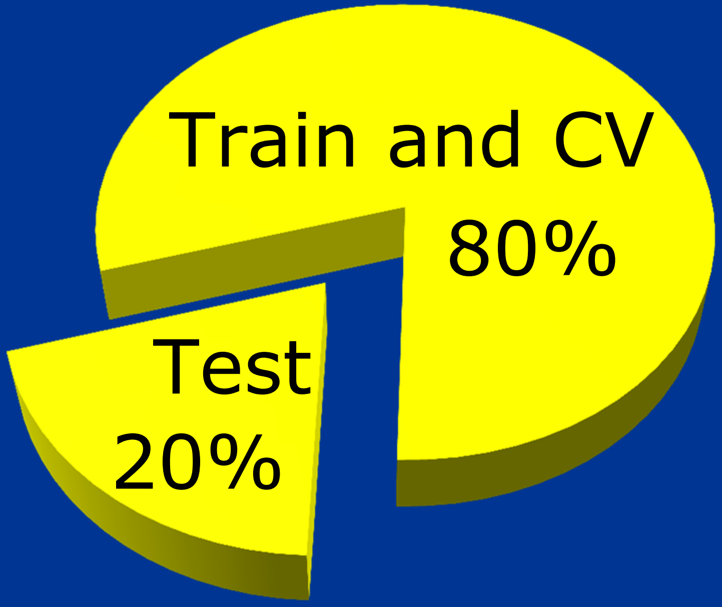
Train and cross-validate models here

Train and CV 80%

Test 20%

Test models here

Goal: Max. Prediction Accuracy



Using the training/CV set, build a model that maximizes the fraction of correctly predicted outcomes in the testing set (prediction accuracy).

Equivalently, minimize the testing set misclassification error.

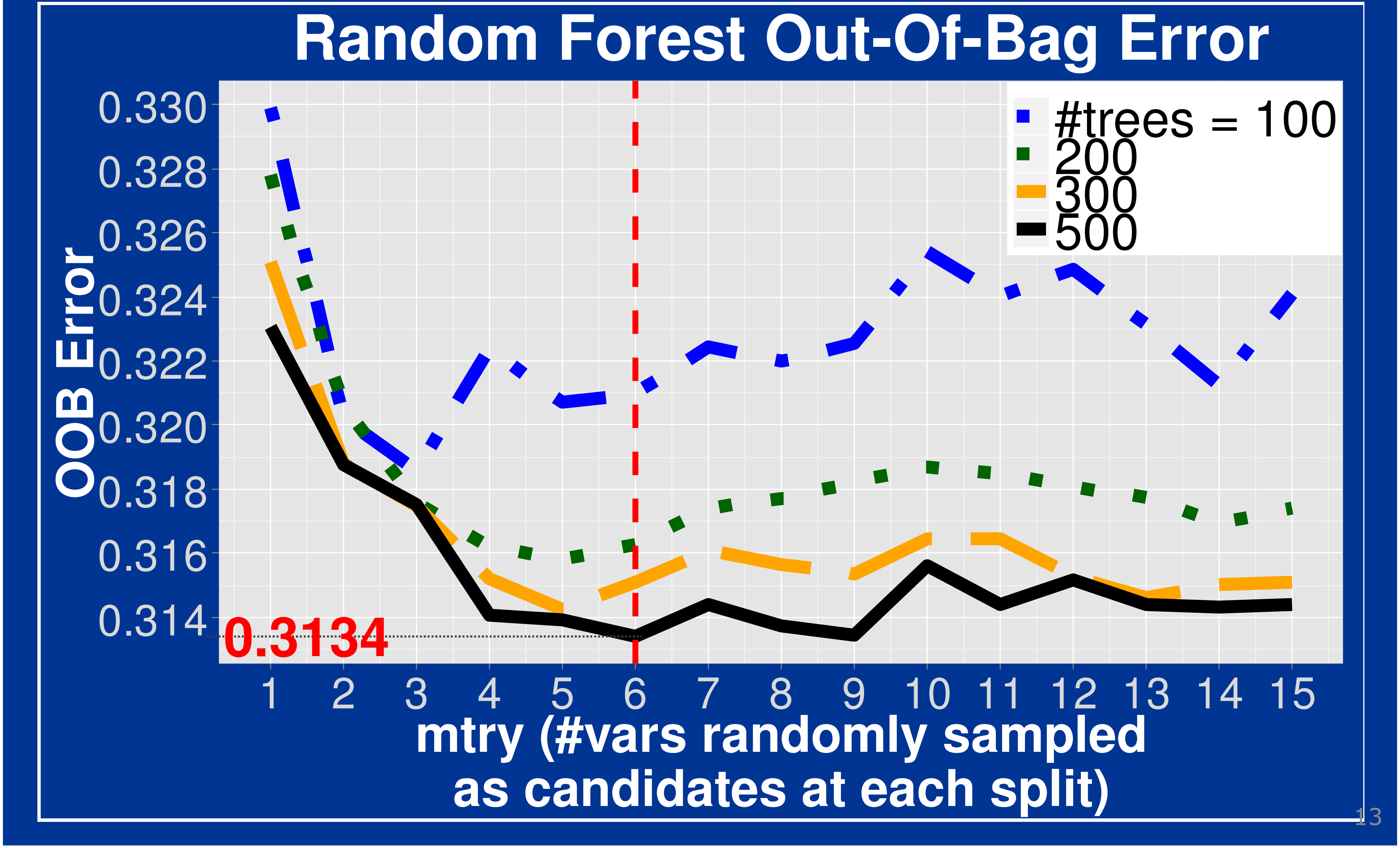
Methods Used

- Random Forest
- Logistic regression
- Regularized logistic regression
  - L1-norm regularization (LASSO)
  - L2-norm regularization (Ridge)

Random Forest

- Ensemble learning method
- Grows multiple decision trees
  - Each tree is based on a bootstrap sample of training data (tree bagging)
  - Each tree, at each decision node, picks a certain number of variables at random and splits the data along one of them

Results: Random Forest



Logistic Regression

- $P\{Y = 1|x\} = G(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$
- Maximize log-likelihood:

$$L = \sum_{i=1}^{N} y_i \log G(x_i\beta) + (1 - y_i) \log(1 - G(x_i\beta))$$

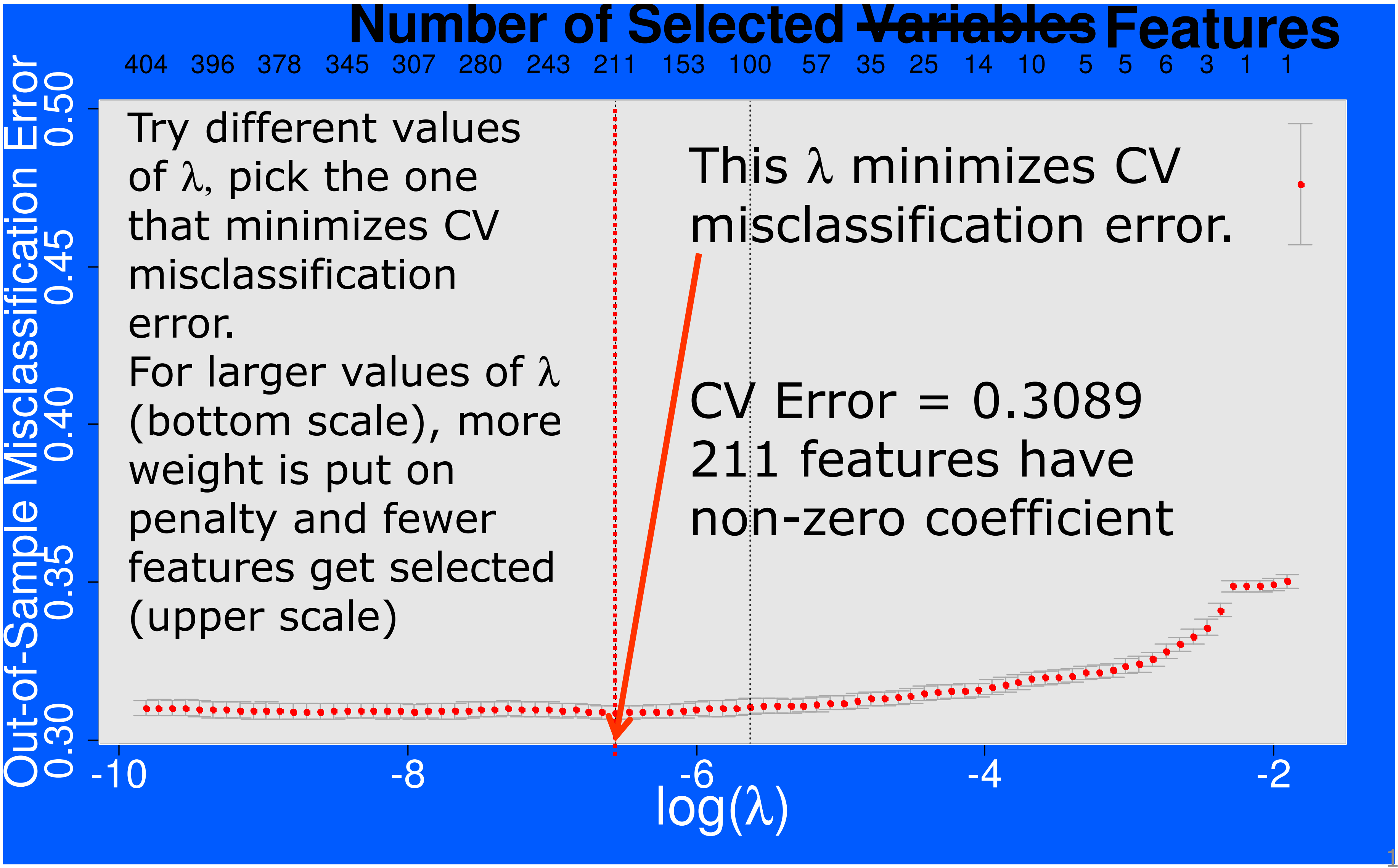
- Fit the entire set of features

Regularized Logistic Regression

$$L = \sum_{i=1}^{N} y_i \log G(x_i\beta) + (1 - y_i)\log(1 - G(x_i\beta)) - penalty$$

- L1-norm penalty: $\lambda\sum_{i=0}^{n}|\beta_i|$ (LASSO)
  - Tends to zero out unimportant features
  - Useful for feature selection
- L2-norm penalty: $\lambda\sum_{i=0}^{n}\beta_i^2$ (Ridge regression)
  - All features get non-zero coefficient
- I use LASSO with 10-fold cross-validation to find optimal λ

LASSO Cross-Validation (CV)



Try different values of λ, pick the one that minimizes CV misclassification error.

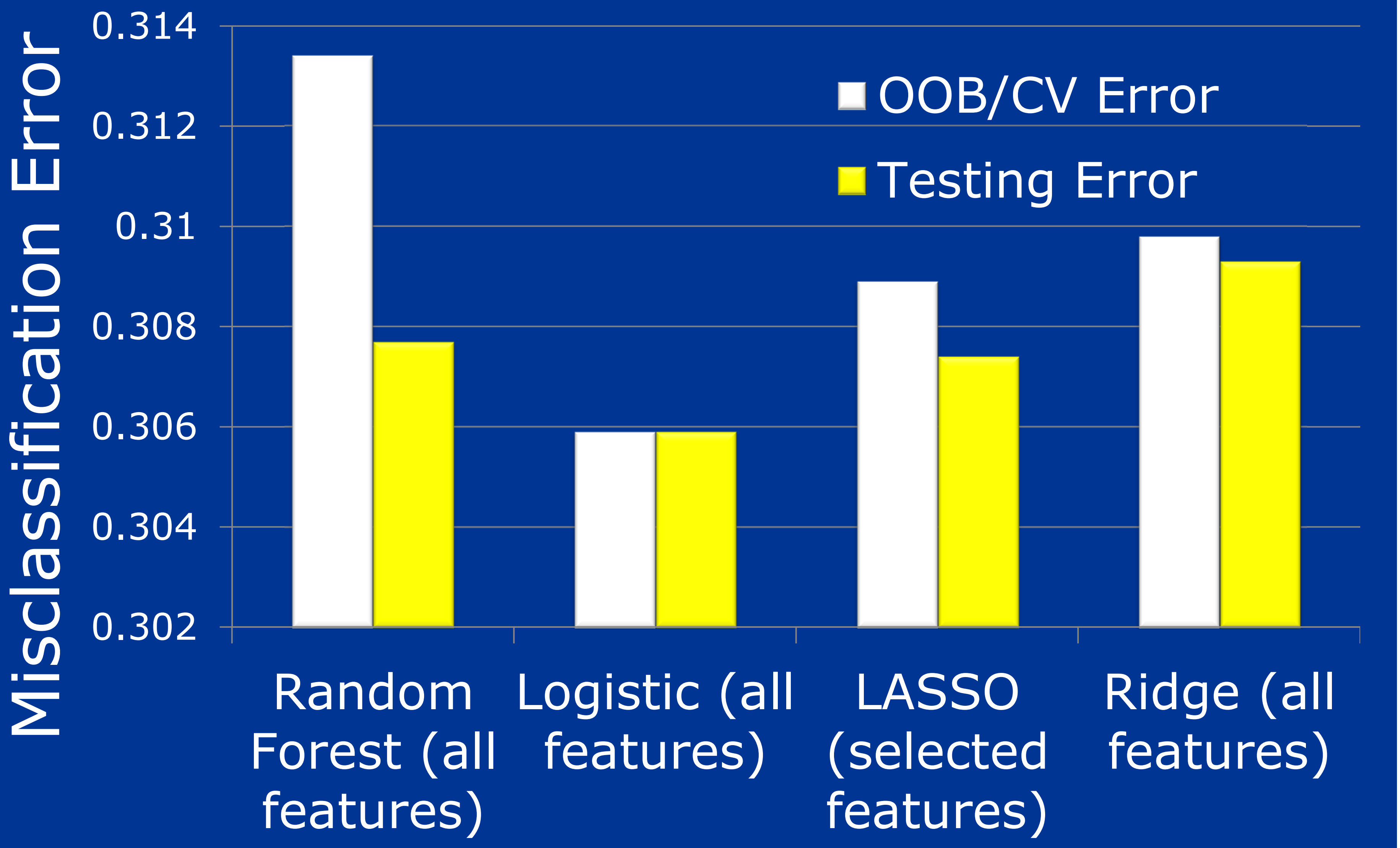
For larger values of λ (bottom scale), more weight is put on penalty and fewer features get selected (upper scale)

This λ minimizes CV misclassification error.

CV Error = 0.3089

211 features have non-zero coefficient

Model Comparison



Best Prediction Model?

- Comparison metric: misclassification error on the held-out testing set (20% of all the data with known outcome)
- Among the four models, logistic regression trained on all features wins by a small margin, with the testing error of 0.3059.

Overfitting?

Q: Could reducing some features *significantly* improve model performance on the held-out testing set?

A: Unlikely. Flat left tail of the LASSO cross-validation plot suggests no overfitting. Moreover, logistic regression trained on the 211 (“best”) features selected by LASSO yields slightly *worse* performance on the testing set (testing error rises from 0.3059 to 0.307).

Feature Selection

- What’s the best parsimonious model?
  - Small subset of the best predictors?
- Missing data affect feature selection
  - Truly important features may be left out because of too many missing values
- How to select features?
  - Using models
  - Expert knowledge

RF-Based Feature Selection

- Random Forest can rank features by importance
- Use it to repeatedly shrink the set of features

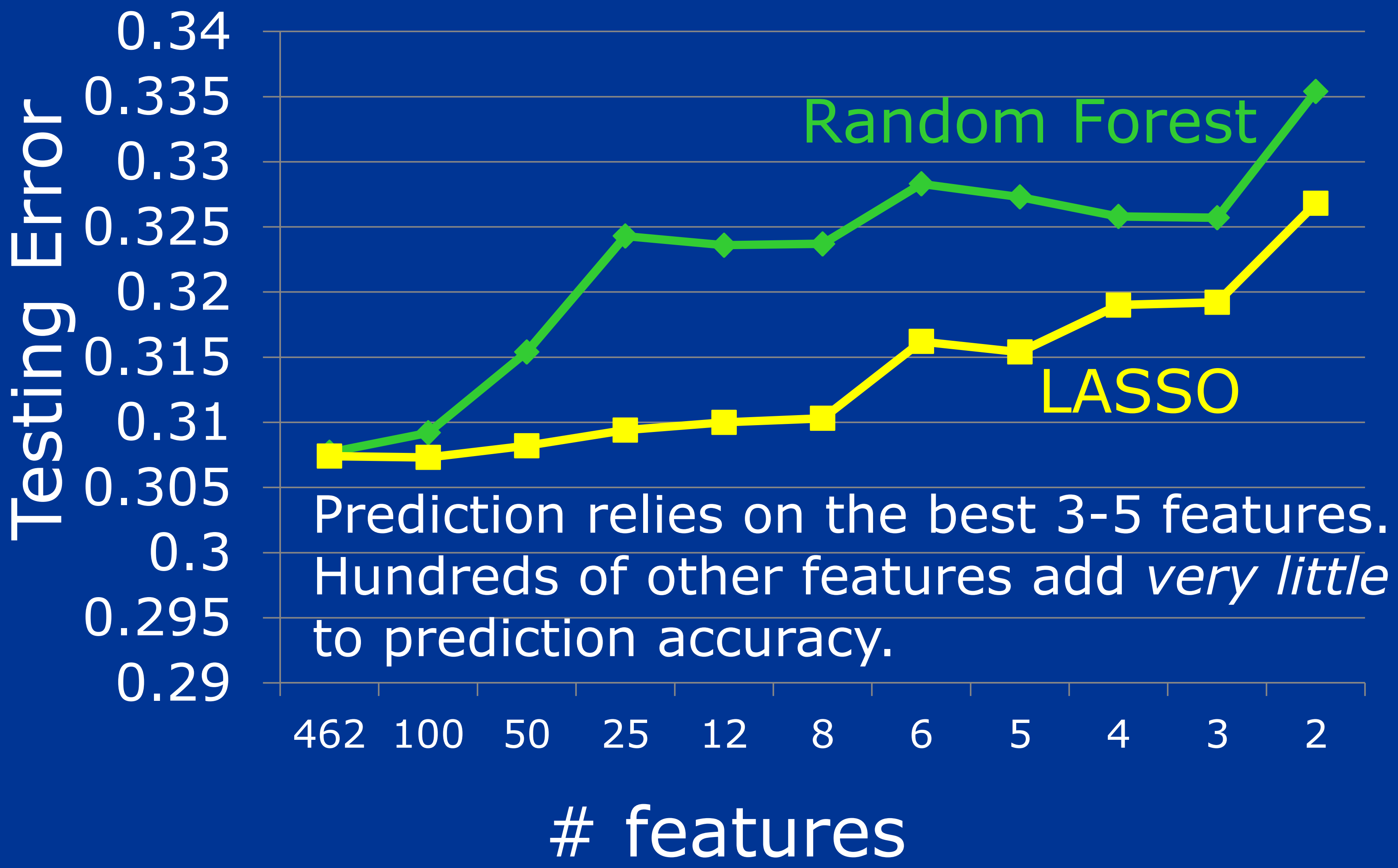
# feat	OOB Err	Test Err
462	0.3134	0.3077
100	0.3147	0.3092
50	0.3189	0.3154
25	0.3255	0.3243
12	0.3353	0.3236
8	0.3255	0.3237
6	0.3300	0.3283
5	0.3288	0.3273
4	0.3290	0.3258
3	0.3296	0.3257
2	0.3382	0.3354

Same performance!

3 Best Predictors:

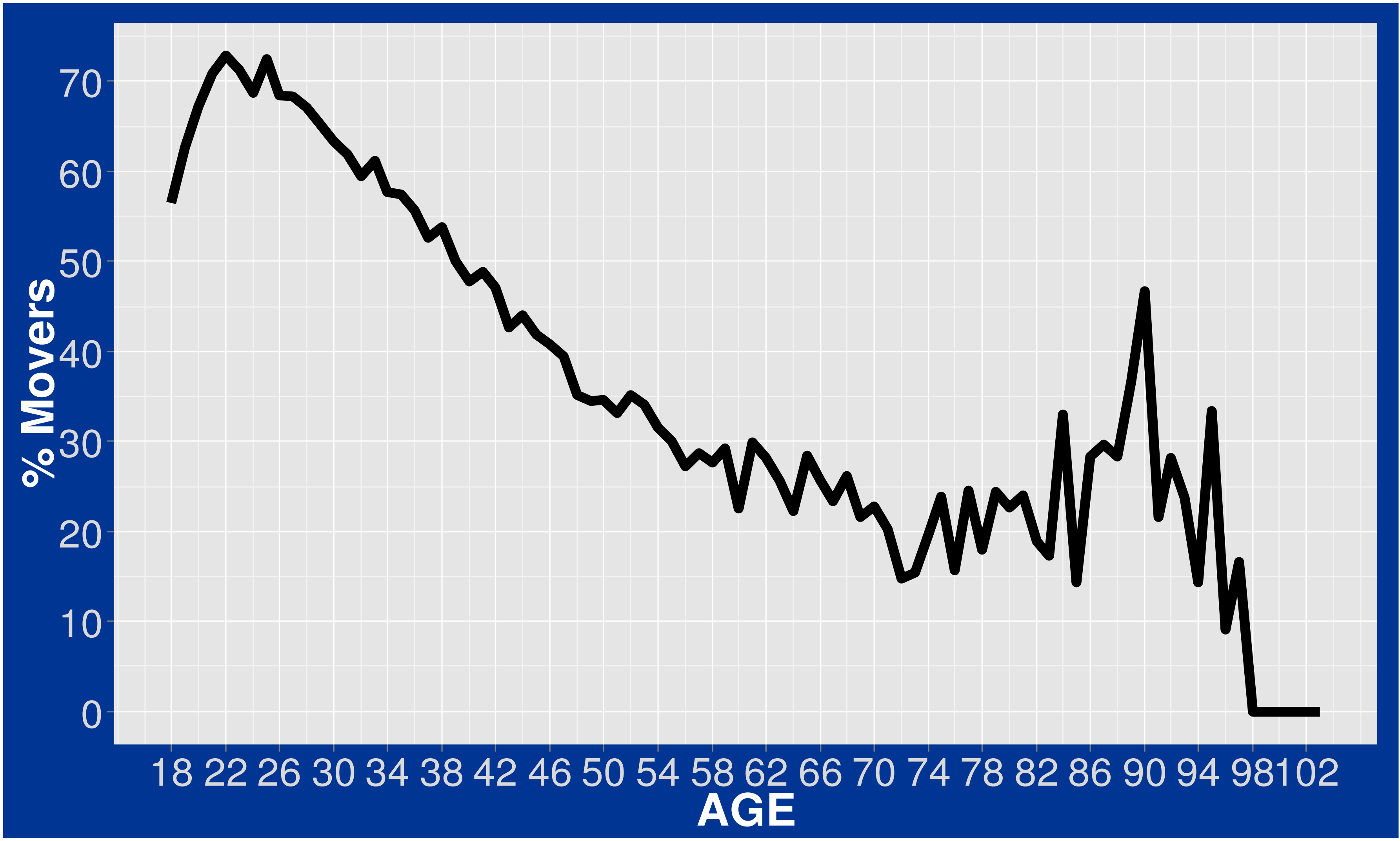
1) Age (age alone yields error 0.35!)
2) Military status
3) "Housing classification"

Feature Selection: RF & LASSO

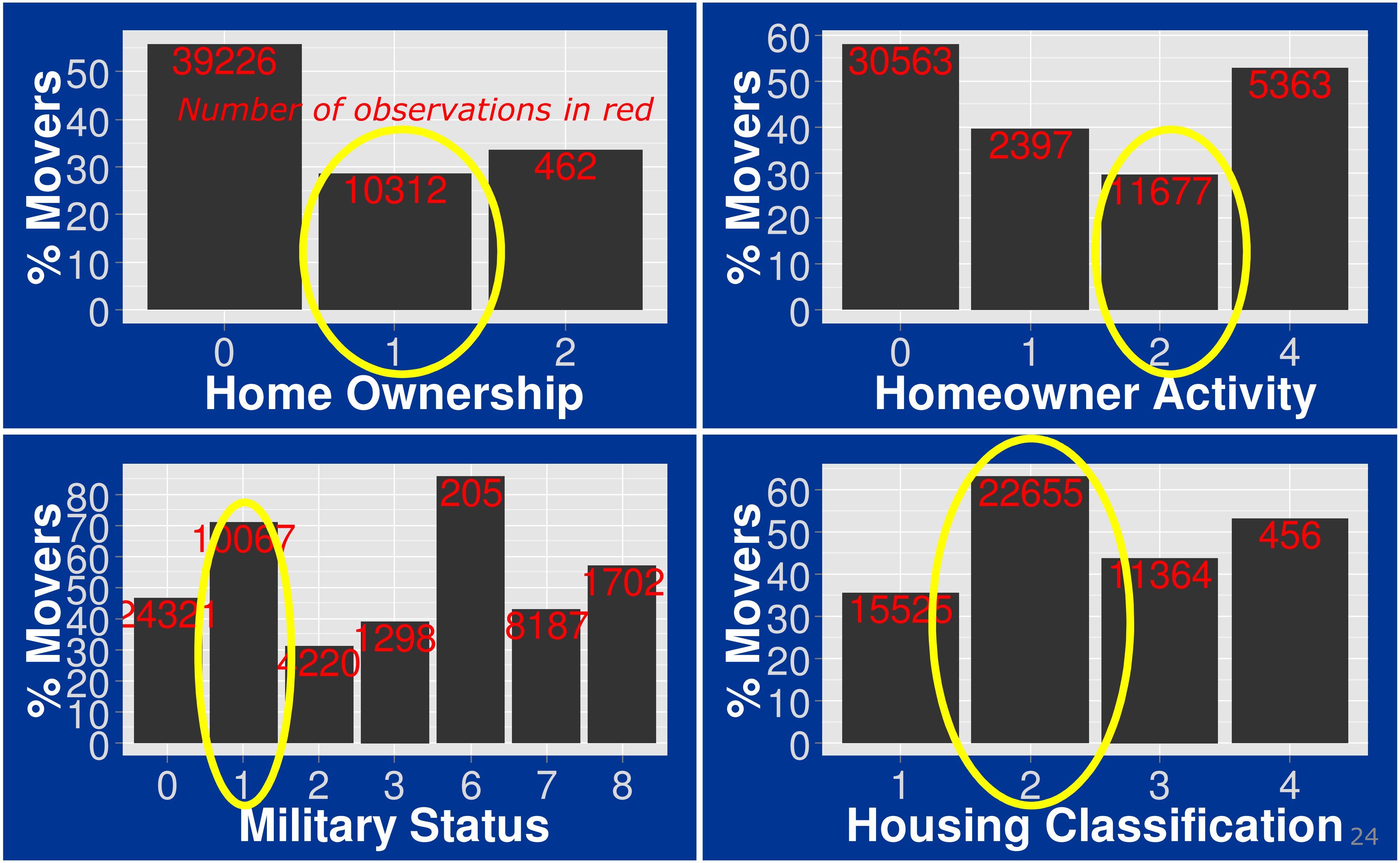


Prediction relies on the best 3-5 features. Hundreds of other features add *very little* to prediction accuracy.

Age Is The Single Most Important Feature



Other 4 Important Features


Number of observations in red

Feature Selection Using Expert Knowledge

What is likely to predict moving?

- Age
  - Younger people move more often
- Military people move all the time
  - Also, Federal law (SCRA Act) allows service members to terminate a rental agreement without penalty if called to active duty or under some other circumstances
- Renters move more often than homeowners
  - It’s easier to move when you rent rather than own a home
- Geography
  - Does location matter for the probability of moving? Urban/rural area? City size?
  - Local unemployment rate?

Why Could Location Matter?

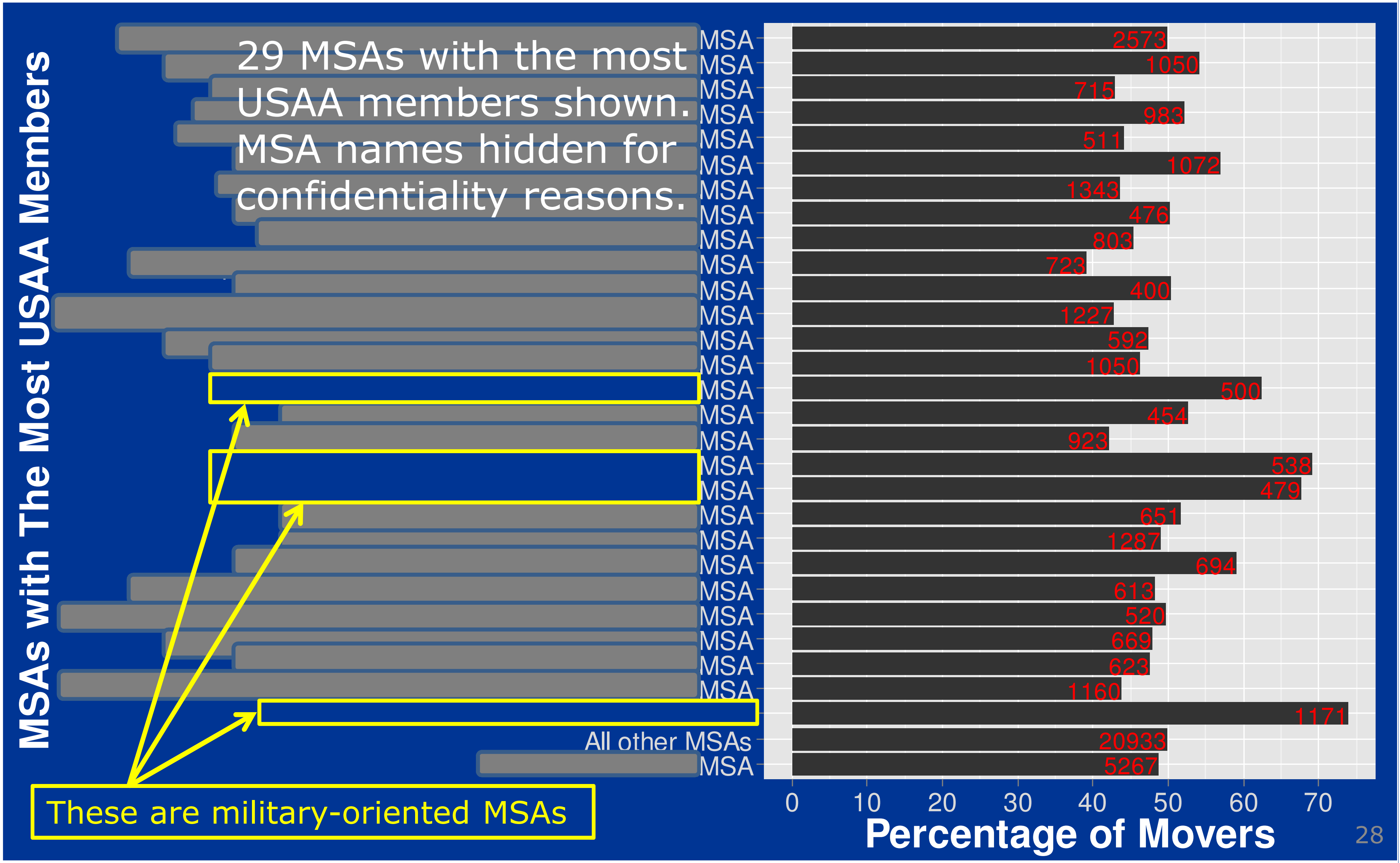
- City size: a job change is more likely to result in a longer commute in a big city than in a small one, other things equal
  - Hypothesis: more moving in bigger cities?
- Local (city-level) unemployment rate (UR) may have two opposite effects on moving: higher local UR 1) increases push migration (moving to a different city); 2) reduces moving within a city due to fewer job changes
  - Hypothesis: local UR may have a non-zero effect on the probability of moving
- Any idiosyncratic unobserved differences across regions could affect the propensity to move

New Data Acquisition

- Original dataset has state and ZIP code but no city or MSA (metropolitan statistical area) variable. I found data on ZIP-MSA pairs, matched ZIP codes in the original dataset to MSAs, and created MSA dummy variables
  - Only 102 observations (out of 50,000) had no matching MSA
- I found data on MSA population size and unemployment rate (UR) at the MSA level in 2013
  - 6,738 observations had no matching UR
  - I imputed missing UR values with median UR

Data Sources: http://www.huduser.org/portal/datasets/usps_crosswalk.html and http://www.bls.gov/lau/lamtrk13.htm

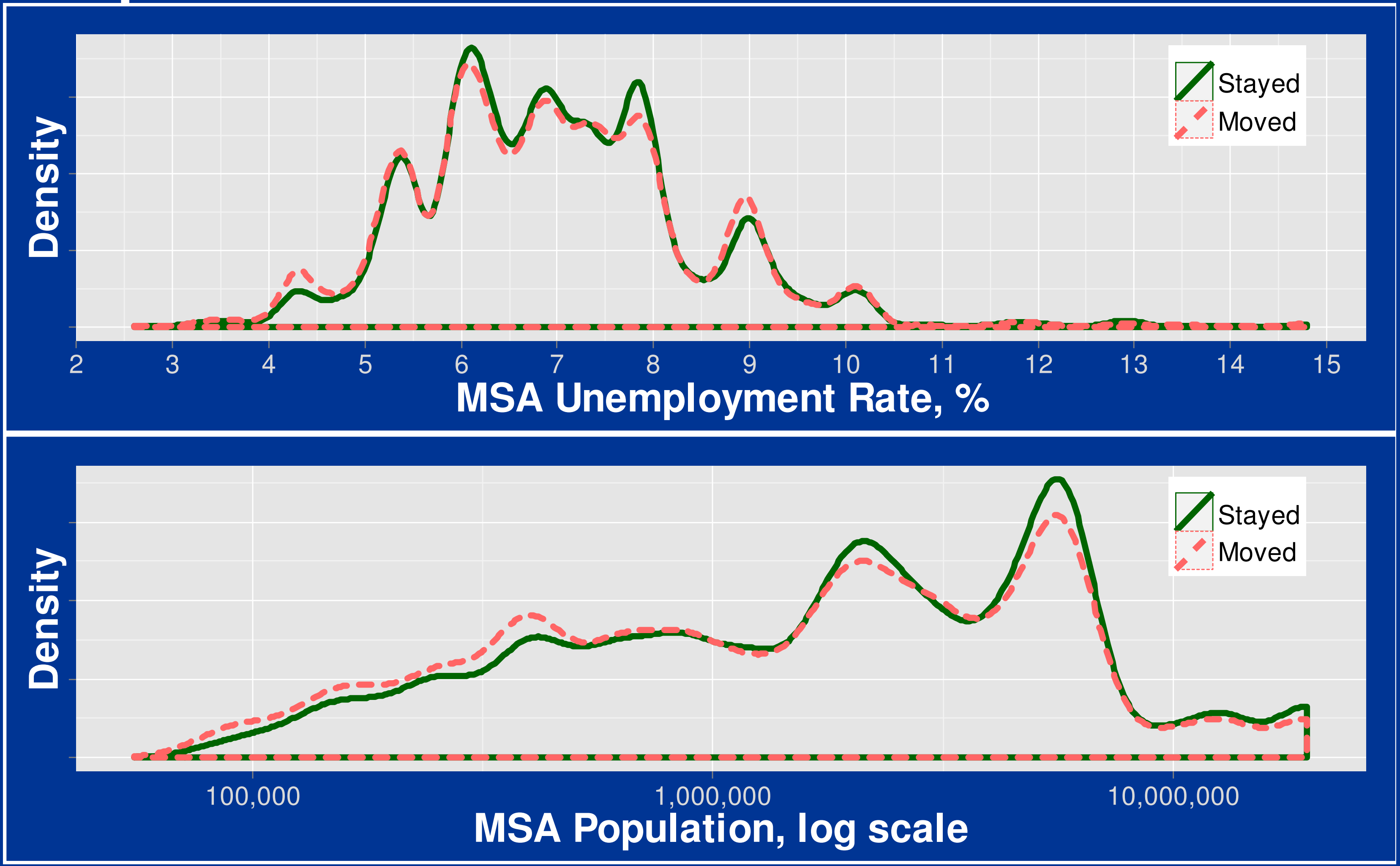
More Moving in Military-Oriented MSAs and Not Much Variation Elsewhere



29 MSAs with the most USAA members shown. MSA names hidden for confidentiality reasons.

These are military-oriented MSAs

MSA Unemployment Rate and Population Size Don’t Matter?



Distributions of MSA unemployment rate for movers (dashed red) and stayers (solid green) in the training dataset look alike (top panel). Same for MSA population size (bottom panel). These are the signs that MSA unemployment rate and population probably don’t matter for the propensity to move.

Testing Statistical Significance

Model: $P\{Y = 1|x\} = G(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$

where $x\beta = \beta_0 + \beta_1 UR + \beta_2 Population +$

$\beta_3 MSA1 + \beta_4 MSA2 + \ldots + \beta_{31} MSA29 +$

betas*(original 462 features)

$MSAi$ is the dummy variable for MSA *i.*

29 MSAs with the most USAA members are included (out of 375 MSAs represented in the data).

Test statistical significance of new features:

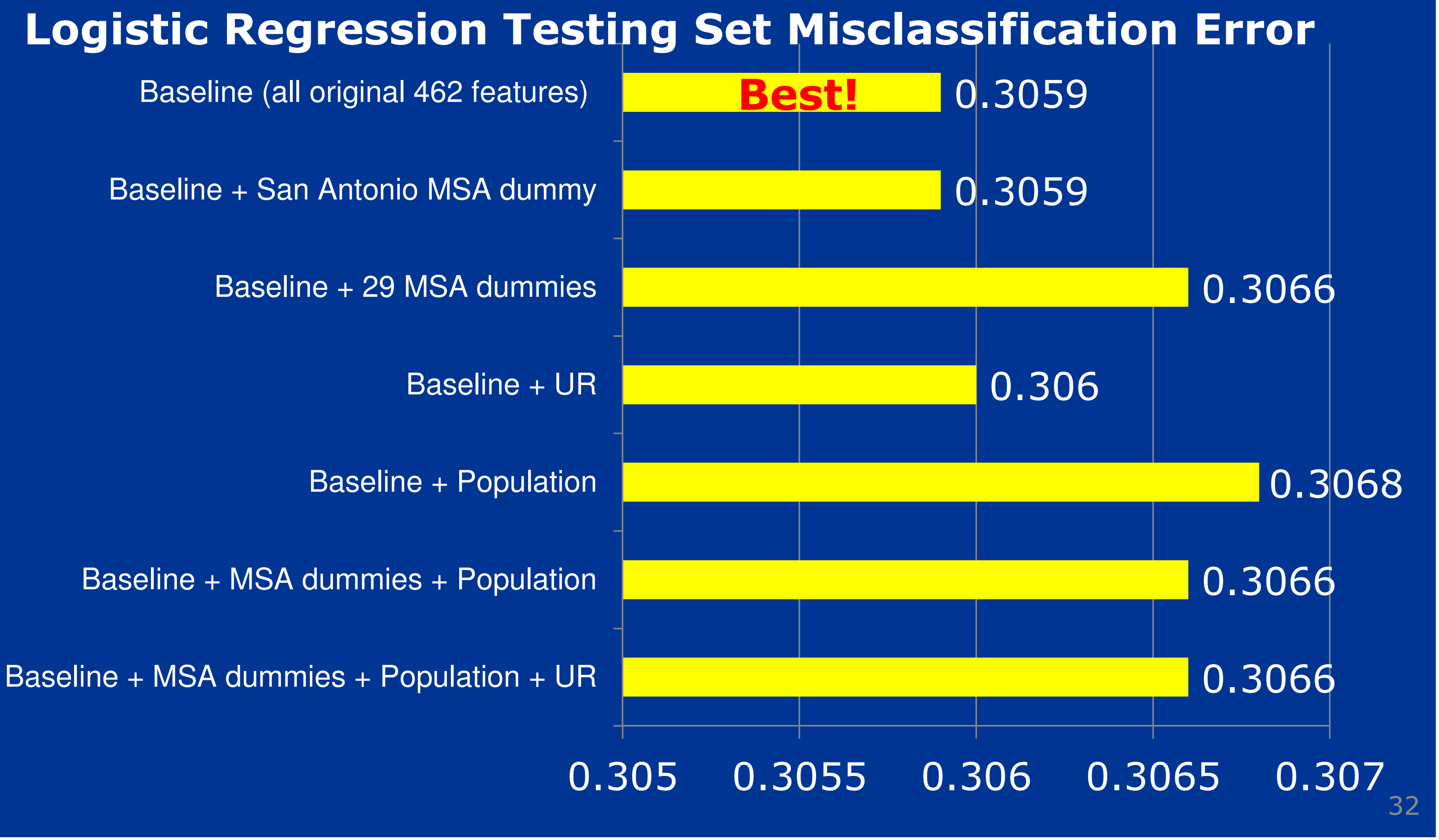
Null hypotheses H_0: $\beta_j = 0$ for $j = 1,\ldots,31$.

Test Results

Feature	Beta Estimate	Std Error	p-value
462 original features	...	...	...
MSA Unemployment Rate	0.0014	0.0111	0.8993
MSA Population	-1.82e-08	2.50e-08	0.4661
MSA1	0.5271	0.6262	0.3999
MSA2	0.0421	0.1603	0.7928
MSA3	-0.0692	0.1268	0.5855
MSA4	0.0117	0.1597	0.9417
MSA5	0.2275	0.2121	0.2835
MSA6	0.3145	0.2531	0.2140
MSA7	-0.1171	0.1668	0.4828
MSA8	0.0691	0.1765	0.6953
MSA9	0.1370	0.1736	0.4299
MSA10	0.1322	0.1372	0.3356
MSA11	0.5448	0.3673	0.1381
MSA12	-0.2761	0.1721	0.1086
MSA13	0.0048	0.1358	0.9718
MSA14	-0.1842	0.1339	0.1689
MSA15	0.2680	0.3104	0.3881
MSA16	0.0448	0.1728	0.7953
MSA17	0.1906	0.4797	0.6911
MSA18	0.7637	0.4790	0.1109
MSA19	0.1031	0.1414	0.4662
MSA20	-0.0676	0.1728	0.6957
MSA21	-0.0751	0.1719	0.6621
MSA22	0.1173	0.1533	0.4444
San Antonio, TX MSA	-0.3672	0.1069	**0.0006*****
MSA24	0.1695	0.1154	0.1418
MSA26	0.0598	0.1549	0.6996
MSA26	0.1247	0.1441	0.3869
MSA27	-0.1969	0.1246	0.1139
MSA28	-0.0750	0.1153	0.5155
MSA29	0.1961	0.1516	0.1958

- These new features are **not** statistically significant at 10% level:
  - MSA unemployment rate
  - MSA population
  - MSA dummies, except for San Antonio, TX
- Insignificance of the dummies for military-oriented MSAs can be explained by high correlation with other features (such as military status)

Adding New Features Does NOT Reduce Prediction Error



Conclusion

- I used my machine learning and data analysis skills to attempt a real business challenge
- Age is by far the best predictor of moving
- 5 best model-selected predictors are consistent with intuition
- Potentially meaningful additional features, such as MSA, population, or local unemployment rate actually do not improve prediction accuracy