Defense Outline

1. Thesis is a solution to problem
    1. too much data, not enough personnel, time, or resources
        1. Further, IV collection is expensive
    2. Best Subset With Validation Algorithm (BeSiVa)
    3. BeSiVa created to help political campaigns
    4. Explain circumstances that led to development
    5. Describe how it works
    6. compare to preexisting variable selection methods
2. Started with my hiring in summer 2013
    1. Hired by Activate as Account Executive
    2. Brought on to develop new means of analyzing campaign data
        1. had several modeling projects for different races
        2. Same problem kept coming up
        3. Large sets of independent variables
        4. glm works, but no way to sift through variables and find relevant ones
        5. different variables for New York City v Ohio
        6. Need to make a prediction of support for each voter
    3. Came to a head with Wakefield campaign for KS second House district
    4. Solved the problem by creating a new Algorithm, BeSiVa, which allows for prediction
3. An Overview of the Independent Variables
    1. Wakefield data was representative
    2. 313 columns initially
    3. Examples of IVs
        1. What party are they and did they vote?
        2. Do people have home mortgages or a car lease?
        3. Do they smoke heavily or subscribe to hunting magazines?
    4. Trimmed down in ad hoc fashion
    5. Even so, 184 IVs remain
    6. contained missing data
        1. brought us down to 185 complete observations from 1,355

4. The Dependent Variable
    1. survey data, collected by Wakefield campaign volunteers
    2. Surveying Methodology (from appendix)
        1. how a predictive dialer works (brief)
    3. Question types
        1. Survey had 3 different question types
            1. Warmup, Issue, Identification,
                1. ID determines candidate support/focus of modeling
    4. 1-10 scale, which became problematic (Show figure 1)
        1. data is highly polarized
            1. 65% 1 or 10, 10% undecided
        2. suggested potential truncation, mention methods to deal
        3. client's desires led me to dichotomize DV
5. The 3 preexisting methodological options for variable selection
    1. **Stepwise and subset selection**
        1. Closest relation to method created
        2. Create sets of models and select 'best' one according to criterion
            1. Best: Creates all models
            2. Forward: start w/1, then add based on individual criteria (i.e. low p vals)
            3. Backward: start with all, then remove based on individual criterion
        3. Problems, as pointed out by Biostatastician Frank Harrell
            1. Biased R^2 to be high
            2. Multicollinearity is nontrivial
    2. **Penalized regression**
        1. Similar to OLS and GLM
        2. 1 key difference: Not unbiased
            1. adds penalty, allowing for IV removal
        3. No sensible way to deal with missing data in IVs
    3. **'Modern' Methods of Adaboost.M1 and random forest**
        1. Random forest: Divides dv into predictions based on divisions it sees in Ivs
        2. Adaboost.m1, classifies, and improves classifier
        3. Each Can deal with missing data
            1. Surrogate splits

    4. results were unimpressive at predicting/interpreting
        1. AdaBoost used a lot of IVS and was hard to interpret
        2. Random forest was similar w/low pred accuracy
6. A Description of the BeSiVa Algorithm
    1. at heart, way of determining Ivs for glm regression
    2. uses logistic regression for campaign data
    3. Go through algorithm.
7. Limitations of BeSiVa's Initial Formulation
    1. Missing data may bias results
    2. Missings may cause inaccurate predictions (if the subset condition occurs)
    3. No way to deal w/situation where two variables are contributing but one isn't
8. Comparison and Results
    1. Bootstrapped the data
    2. 50 runs, taking different subsets for training and test data for each
    3. Put PCP for each method into categorization and regressed against category
        1. (show table 2 and density plot, and discuss)
9. Future Considerations
    1. Make this a function
    2. increase types of glm
    3. add things besides AIVs to selection format (i.e. priority ranking, go through text)
10. To reiterate,