# Sensitivity of Jury Trial Outcomes to Trial Factors

Joseph M. Abruzzo
Stanford University

This paper presents a novel machine learning approach to feature impact analysis in the context of American criminal jury trials. Specifically, four data mining algorithms are used to rank nineteen characteristics of jury trial procedure according to their influence on trial outcomes. The algorithms are implemented using various packages in the statistical computing language R. Implications, limitations, and directions for future research are discussed.

## Introduction

Statistical analysis of criminal trial jury data sets is not a new idea. The University of Chicago Jury Project was the initial large-scale effort both to gather jury trial data and to relate important social variables to jury trial outcomes. The Project was successful in its latter goal, generating an enormous reservoir of data that was used by subsequent researchers throughout the 1960s (Newman & Remington, 1966; Broeder, 1964, 1965; Okun, 1967; Nagel, 1966).

Jury research since this early era has focused primarily on the effects of juror characteristics on trial outcomes, and in the social sciences has revolved around behavioral studies of both real and mock juries. Such studies generally display interaction effects between the defendant's demographic characteristics and those of the jurors, with the upshot of the body of work being that interaction effects for race essentially bias jury trials against Black defendants (Perez, Hosch, Ponder, & Trejo, 1993; Klein & Klastorin, 1999; Sommers, 2006, 2007; Sommers & Norton, 2007). Research from the legal domain has gone so far as to create stochastic simulations of these effects, some of which display robust predictive power (Kerr, Stasser, & Davis, 1979; Boster, Hunter, & Hale, 1991; Brown, Doyle, et al., 1996; Wittlin, 2016).

With research at the mathematical simulation stage, machine learning-style mining of jury trial data sets has declined in popularity since its debut in the 1950s. Where it does appear, the techniques used rarely extend in complexity beyond simple linear regression (Hannaford-Agor & Hans, 2003; Anwar, Bayer, & Hjalmarsson, 2010, 2013). While not necessarily a mistake, simple linear regression is highly biased because of its intentional simplicity, and therefore seems a rather blunt instrument with which to conduct research on an intricate social phenomenon.

The research presented here extends previous work in two ways. First, it updates the data mining approach to jury research to include more advanced statistical techniques, and second, it explores the possibility of influences on jury trial outcomes that are not explicitly related to juror demographic characteristics. This paper presents a feature impact analysis of trial procedure characteristics on trial outcomes using four machine learning algorithms: a logistic regression with LASSO, a penalized support vector machine, a random forest, and a neural network.

## Data Set

The data set for this project is from research conducted in 2003 on hung juries and jury nullification (Hannaford-Agor & Hans, 2003). Despite the authors' restriction of their analysis to juror, defendant, charge, and case characteristics, the scope of the project's data included over thirty features of trial procedure. The data set has been made available to academic researchers through the University of Michigan's National Archive of Criminal Justice Data (*Evaluation of hung juries in Bronx County, New York, Los Angeles County, California, Maricopa County, Arizona, and Washington, DC, 2000-2001 (ICPSR version) [Data File]*, 2003). It contains data from 351 criminal jury trials collected by surveying judges, attorneys, and jurors in Bronx County, New York, Los Angeles County, California, Maricopa County, Arizona, and Washington, D.C.

### Missing Data and Data Imputation

Because the data set was gathered via paper survey, it contains missing elements where participants failed to respond to researchers' questions. While most of the features were complete enough to be salvageable by data imputation, other features were missing enough observations to render them useless. It was decided that features missing more than 20% of their observations would be discarded from the data set. After this process was completed, nineteen features remained. These are listed below, along with their designations in the final data set.

1. The length of the voire-dire process in hours (`voireDireLength`).
2. Whether or not a paper questionnaire was used in the voire-dire process (`voireDireQuestion`).
3. The number of jurors struck by the prosecution (`struckPros`).

4. The number of jurors struck by the defense (`struckDefen`).
5. Whether or not the jury was selected anonymously (`anonJury`).
6. The length of the trial in hours (`trialLength`).
7. The number of witnesses for the prosecution (`witnessPros`).
8. The number of expert witnesses for the prosecution (`expWitnessPros`).
9. The number of exhibits used by the prosecution (`prosExhibit`).
10. The number of witnesses for the defense (`witnessDefen`).
11. The number of expert witnesses for the defense (`expWitnessDefen`).
12. The number of exhibits used by the defense (`prosExhibit`).
13. Whether or not jurors were permitted to take notes (`noteTake`).
14. Whether or not jurors were provided notebooks with which to take notes (`notebook`).
15. Whether or not jurors were permitted to submit questions to witnesses (`jurySubQuest`).
16. Whether or not jurors were aware of the sentencing possibilities for the defendant resulting from a guilty verdict (`juryAwareSentence`).
17. Whether or not conduct guidelines were given for jury deliberations (`guidedConduct`).
18. The number of jurors for the trial (`numJurors`).
19. The length of the jury deliberation procedure in hours (`delTime`).

Missing observations for the above features were imputed using the Predictive Mean Matching (PMM) algorithm. This algorithm is outlined below.

1. Linearly regress the features with missing observations $X_M$ on the features with no missing data $X_{\sim M}$, generating a vector of coefficients $\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$.
2. Select a new vector of coefficients $\theta^*$ from the distribution of $\theta$.
3. Produce estimates of the missing observations $x_M^{(i)} \in X_M$ using the new coefficients $\theta^*$.
4. For each missing observation $x_M^{(i)}$, choose the observation $x_M^{(j)} \in X_M$ with the observed value closest to the estimate calculated in (3).
5. For each missing observation $x_M^{(i)}$, set $x_M^{(i)} := x_M^{(j)}$.

The R package `mice` provides a convenient implementation of the PMM algorithm. The package was used to generate an imputed data set that was used for the remainder of the project.

Note that it is generally ill-advised to impute categorical variables, as the last two steps of the PMM algorithm clearly do not make sense when applied to variables on discrete scales. As such, only continuous variables are imputed.

### Dependant Variable

The outcomes of jury trials are not necessarily binary. Juries sometimes fail to reach a decision for a myriad of reasons. A defendant may be charged with more than one crime, and may be convicted on some of her charges and acquitted on others. Moreover, multiple defendants may be tried at the same time. To simplify the outcome measure, this analysis only considers cases with exactly one defendant. To further simplify the outcome measure, each case is classified with the value 1 if the defendant was convicted on at least one charge and 0 otherwise.

After clearing cases with multiple defendants and cases where verdict data was not provided, the data set was left with 315 cases that robustly represented all four possible trial locations. These cases were well-balanced in terms of the outcome variable just described, with 61.9% of the cases labeled 1.

## Analysis and Results

Four different approaches are used to rank the impact of the nineteen features, with the supposition being that overlap among the separate resultant rankings would indicate especially important overall impact. Only the top five most impactful features from each ranking method are presented.

### Logistic classification with LASSO

The first approach uses logistic classification in combination with the LASSO method for variable selection and regularization to first eliminate variables that do not increase classification accuracy and then rank features in terms of the absolute value of their coefficients, where a greater absolute coefficient value is treated as an indicator of greater impact. The R package `elnet` provides a convenient implementation of LASSO.

$k$-fold cross validation with $k = 5$ is used to determine an optimal value for the LASSO regularization parameter $\lambda$ with respect to misclassification error. The optimal value is $\lambda = 0.051$ (See Figure 1), which corresponds to a logistic classifier with `witnessDefen`, `expWitnessPros`, and `jurySubQuest` as features (See Figure 2). The resultant coefficients and corresponding impact ranking are presented in Table 1.

### Penalized Support Vector Machine

The second approach uses a support vector machine with L1-norm regularization to simultaneously select features and weight them. In a similar fashion to the LASSO regularization employed above, 5-fold cross validation is performed to select an optimal value of the SVM's tuning parameter $\epsilon$

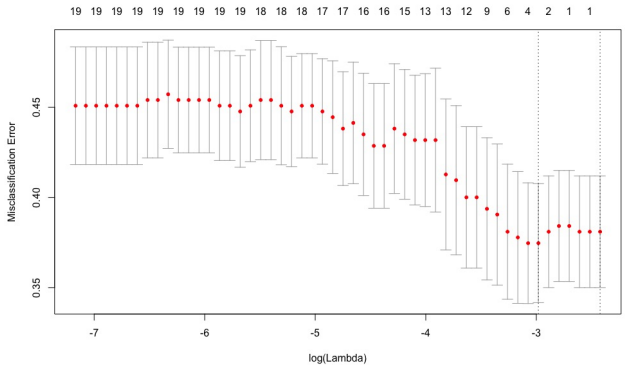


*Figure 1*. Displays the logarithm of the range of LASSO regularization parameters $\lambda$ with respect to their mean 5-fold classification errors. The optimal value of lambda is marked with a vertical dotted line.

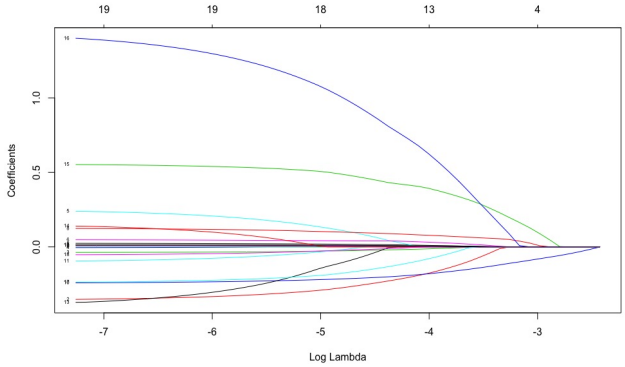


*Figure 2*. Displays the logarithm of the range of LASSO regularization parameters $\lambda$ with respect to their mean 5-fold classification errors. The model corresponding to the optimal value of $\lambda$ can be expressed visually by tracing all of the curves intersecting a vertical line at the optimal $\lambda$.

according to which value produces the smallest average misclassification error. The R package `penalizedSVM` provides an implementation of this technique. Selecting the optimal $\epsilon$, an SVM with 15 features is produced. These are ranked according to the absolute values of their fitted weights. The weights and corresponding impact rankings are presented in Table 1.

## Random Forest

The third approach uses a random forest classifier combined with a "leave one out at a time (OAT)"-style sensitivity analysis. The R package `randomForest` provides an implementation of a basic random forest classifier with a function that allows for $k$-fold cross validation with the variables left out OAT-style. Choosing $k = 5$, features are then ranked in impact according to the mean misclassification error when

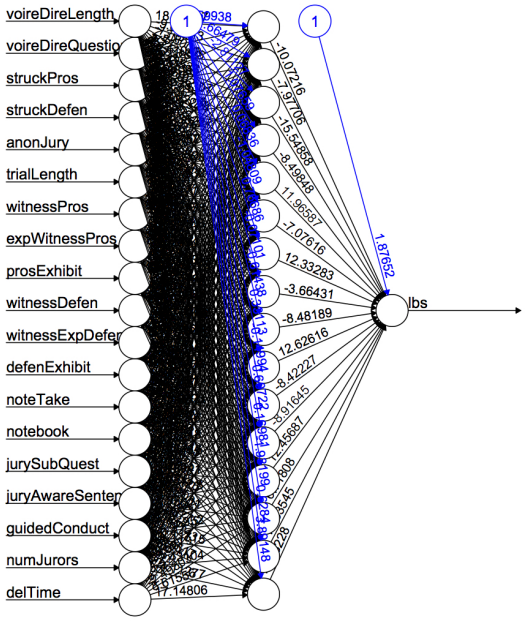


*Figure 3*. A visual representation of the RNN used as the fourth method of assessing feature impact. The output variable `lbs` is a re-coding of the outcome variable `outcome` such that its values are in $\{-1, 1\}$ rather than in $\{0, 1\}$.

they are left out. The top five resultant mean misclassification errors and corresponding impact rankings are presented in Table 1.

## Neural Network

The fourth approach uses a recurrent neural network with one hidden layer containing 16 hidden nodes to rate feature impact (See Figure 3). The weights of the features in the trained neural net are taken as an index of the features' relative impact. The weights of the features and the corresponding impact ranks are presented in Table 1.

## Discussion

From Table 1, five features made appearances in more than one ranking scheme: `jurySubQuest`, `witnessDefen`, `struckPros`, `sruckDefen`, and `numJurors`. Of these, `witnessDefen` most consistently ranks the highest, so it may be declared the most impactful feature in the data set. This result is interesting, and several interpretations may be offered to explain it.

It is possible that jurors place high weight on testimony on behalf of the defendant. If witnesses for the defense are more likely to attest to the character of the defendant, jurors might be less likely to prefer a conviction going into deliberations. It is also possible that witnesses for the defense provide unusually strong evidence for the exoneration of the defendant

Table 1
*Output of ranking methods and corresponding feature impact rankings*

| Rank | Logistic & LASSO | Penalized SVM | Random forest & OAT | Neural network |
|---|---|---|---|---|
| 1 | `jurySubQuest` [0.0897] | `witnessDefen` [-0.2106] | `witnessDefen` [0.0154] | `numJurors` [120.72] |
| 2 | `witnessDefen` [-0.0794] | `juryAwareSentence` [0.1019] | `defenExhibit` [0.0124] | `prosExhibit` [72.24] |
| 3 | `expWitnessPros` [0.0081] | `guidedConduct` [-0.0678] | `jurySubQuest` [0.0077] | `witnessPros` [66.23] |
| 4 | N/A | `trialLength` [0.0546] | `struckDefen` [0.0048] | `struckDefen` [56.90] |
| 5 | N/A | `struckPros` [-0.0308] | `numJurors` [0.0036] | `witnessDefen` [53.34] |

by providing alibis, pointing to other suspects, etc.

Equally interesting are the features that do not register as impactful. While `witnessDefen` is the most impactful feature, some seemingly related or complementary features, such as `witnessPros`, `expWitnessDefen`, and `expWitnessPros`, do not rank in the top five. Furthermore, features of the trial that should be the most influential according to their theoretical relevance to evidence strength, such as `prosExhibit` and `defenExhibit`, also do not make the top five. Note also that each of the top five most influential features have either to do with which jurors end up on the seated jury (i.e. `struckPros`, `struckDefen`, and `numJurors`) or how those jurors are then presented evidence (i.e. `witnessDefen` and `jurySubQuest`). One might take this to mean that who is hearing the evidence and how that evidence is presented to them has more leverage on the ultimate outcome of jury trials than that evidence itself.

Because this research is limited in scope and because it does not employ experimental methodology, no explicit claims may be made about the mechanisms by which the imapctful features influence trial outcomes. However, the results of this project do suggest the strong influence of cognitive biases in the jury trial process, which complements the robust and ever-expanding body of research supporting that notion.

## Limitations

This classification problem in this project may be described as a difficult one. The most successful classifier that was tested in this research was the logistic classifier with LASSO, which achieved a misclassification error rate of 37.5%, just 0.6% under the trivial error rate of 38.1%. This was expected from the outset of the project, as no features that have historically been associated with conviction or exoneration in jury trials were included. With this data set, at least, machine learning algorithms have not proved themselves extraordinarily useful for purely predictive purposes. It remains to be seen in subsequent research with more robust data sets and larger feature spaces if the machine learning approach can be used successfully for predicting the outcomes of individual trials. Nevertheless, data mining has certainly demonstrated its usefulness for strictly inferential purposes in this domain.

## Conclusion

This project successfully employed modern machine learning techniques to better understand the impact of previously under-explored features on jury trial outcomes. Its conclusion is that the features that have the most impact are those that are only loosely connected to the strength of the evidence against the defendant. While the approach used here has its limitations, it surely has merit and may be employed productively in future jury research.

## Acknowledgements

Special thanks to Francois Germain and the CS 229 TA team for their advice and feedback throughout this project. Special thanks to Prof. Andrew Ng and Prof. John Duchi for the wonderful class. Special thanks to Dr. Susan McInnes, Lauren Abruzzo, Sara Berg-Love, Dr. Elizabeth Sather, and Erika Taddie-Osborn for their support.

## References

Anwar, S., Bayer, P., & Hjalmarsson, R. (2010). The impact of jury race in criminal trials.

Anwar, S., Bayer, P. J., & Hjalmarsson, R. (2013). The role of age in jury selection and trial outcomes. *Economic Research Initiatives at Duke (ERID) Working Paper*(146).

Boster, F. J., Hunter, J. E., & Hale, J. L. (1991). An information-processing model of jury decision making. *Communication Research*, *18*(4), 524–547.

Broeder, D. W. (1964). Voir dire examinations: An empirical study. *S. Cal. L. Rev.*, *38*, 503.

Broeder, D. W. (1965). The negro in court. *Duke Law Journal*, *1965*(1), 19–31.

Brown, D. C., Doyle, T. K., et al. (1996). A computer simulation model of juror decision making. *Expert Systems With Applications*, *11*(1), 13–28.

*Evaluation of hung juries in bronx county, new york, los angeles county, california, maricopa county, arizona, and washington, dc, 2000-2001 (icpsr version) [data file].* (2003). (Available from http://doi.org/10.3886/ICPSR03689.v1)

Hannaford-Agor, P. L., & Hans, V. P. (2003). Nullification at work-a glimpse from the national center for state courts study of hung juries. *Chi.-Kent L. Rev.*, *78*, 1249.

Kerr, N. L., Stasser, G., & Davis, J. H. (1979). Model testing, model fitting, and social decision schemes. *Organizational Behavior and Human Performance*, *23*(3), 399–410.

Klein, K. S., & Klastorin, T. D. (1999). Do diverse juries aid or impede justice. *Wis. L. Rev.*, 553.

Nagel, S. S. (1966). Disparities in criminal procedure. *UCLA L. Rev.*, *14*, 1272.

Newman, D. J., & Remington, F. J. (1966). *Conviction: The determination of guilt or innocence without trial*. Little, Brown Boston.

Okun, J. (1967). Investigation of jurors by counsel: Its impact on the decisional process. *Geo. Lj*, *56*, 839.

Perez, D. A., Hosch, H. M., Ponder, B., & Trejo, G. C. (1993). Ethnicity of defendants and jurors as influences on jury decisions1. *Journal of Applied Social Psychology*, *23*(15), 1249–1262.

Sommers, S. R. (2006). On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. *Journal of personality and social psychology*, *90*(4), 597.

Sommers, S. R. (2007). Race and the decision making of juries. *Legal and Criminological Psychology*, *12*(2), 171–187.

Sommers, S. R., & Norton, M. I. (2007). Race-based judgments, race-neutral justifications: Experimental examination of peremptory use and the batson challenge procedure. *Law and Human Behavior*, *31*(3), 261–273.

Wittlin, M. (2016). The results of deliberation. *University of New Hampshire Law Review*, *15*(1), 161–225.