Predicting Supreme Court Decision Making

Monica Rajendiran

Executive Summary

The decisions of Supreme Court cases have been analyzed by legal experts, scholars, and hobbyists for decades but recently their decisions are also coming under the lens of applied machine learning and statistical modeling techniques. This is a challenge because the cases that Supreme Court justices get are often unclear and complex in light of Constitutional Law. I explore how to predict Supreme Court Decisions (the binary outcome of whether the Court affirms or reverses the judgement of the lower courts) through historical case data, provided by the Supreme Court Database. This historical case data contains information on every case from 1946 to 2015, containing variables such the date an Oral Argument was held or what issue the case pertained to. I have also added several engineered features to complement this dataset, intending to draw information from outside areas such as the economy (e.g. unemployment rate for every Supreme Court term) and the political atmosphere at the time (e.g. the makeup of republicans and democrats in Congress for every Supreme Court term).

My exploratory analysis of the distributions and variations within the court predictors provided a number of insights into the decision making process of the Supreme Court. For example, I was able to see that the Supreme Court had taken on a decreasing number of cases since the 1980s. This observation was confirmed by external research. I also used basic text analysis and time series analysis to obtain exploratory observations in the data. For example, I analyzed the top 20 words in 2001 - 2002 oral arguments by frequency through a TF-IDF matrix and created a time series object in order to examine the number of cases that occurred each month in each Supreme Court term for any cyclical information.

My primary hypothesis was predicting the Supreme Court's decision to affirm or reverse the judgement of a lower court for a particular case. I developed logistic regression models using hybrid stepwise selection, developed random forest models, and also developed a support vector machine model to test this hypothesis. I used confusion matrices, variable importance plots, 20-fold Cross Validation, Raw Accuracy, and AUC metrics to evaluate the models. I held out 25% of the observations at random as a test set in order to evaluate testing performance. Random Forest performed the best out of Logistic Regression and Support Vector Machines. The Random Forest had an AUC score of .87, the Logistic Regression model had a score of .67, and the SVM had a score of .66. Comparing the Random Forest and SVM, Random Forest had a raw accuracy score of 79.8% while SVM had a raw accuracy score of 64.8%. According to our confusion matrix, our sensitivity for Random Forest is satisfactory (0.6639) so these **results do support our hypothesis** that we can predict whether a Supreme Court rules to affirm or reverse a decision based on historical case information. However, on a non-random train/test split based on chronology, the Random Forest performed poorly (.58 AUC), most likely due to a poor split between the percentage of responses in the data when accounting for chronology.

1. Problem Identification

Descriptive Analysis

The Supreme Court in its current form exists as the pinnacle of federal judicial decision making, the final interpreter of the United States Constitution. It has ultimate appellate jurisdiction, meaning that it has the ability to review and change the decisions made by lower courts. In its usual form, the court consists of a Chief Justice that presides over 8 other justices that are categorized as conservative, liberal, or moderate in their interpretation of the law and Constitution. Justices are nominated by the President and confirmed by the Senate but they are expected to be independent and nonpartisan in their interpretation of the Constitution and their subsequent passing of judgements on cases that bubble up to them from the lower courts [1].

The decisions of Supreme Court cases have been analyzed by legal experts, scholars, and hobbyists for decades, but recently their decisions are also coming under the lens of applied machine learning and statistical modeling techniques. This is a challenge because the cases that Supreme Court justices get are often unclear and complex in light of Constitutional Law. On top of that, the issues are handled by human justices with deep, unique, and often evolving ideologies and interpretations which can be hard to quantify.

Although Supreme Court justices are expected to be independent in their decisions, it is clear from legal and historical analysis that external influences that can potentially be quantified affect the justice's decisions. David W. Rohde and Harold J. Spaeth, in their book "Supreme Court Decision Making", argue that Supreme Court decisions are based on the factors: goals, rules, and situations [2]. Individual justices have personal policy goals that are influenced by their beliefs and values. Justices' decision making based on their policy goals are constrained by the rules of the court, as well as situational factors such as public opinion and political climates. Researchers Casillas, Enns, and Wohlfarth found that public opinion serves as a situational constraint on the decisions that the Court makes [3]. For example, they found that the Court exhibited a roughly 1% increase in the proportion of liberal decision making for every 1% shift in public opinion towards liberal ideologies. Along with personal ideologies, beliefs, and values, one study found that in a substantial number of cases, precedent was an important factor in the decision that an individual judge may make, perhaps more so than personal ideology in some cases [4].

The are many factors that influence Supreme Court decision making creates a large space to analyze whether those factors can lead to predicting decisions. In fact, there are several studies that focus on studying the Court and attempting to predict voting behavior in any given year with a given case. For example, a law professor from the University of Pennsylvania, along with a team of legal scholars and political scientists, undertook a study to predict cases in the 2002 -

2003 Supreme Court term using classification trees based on ideological preferences and past voting history [5]. Another study looked at six decades of Supreme Court votes, variables, and cases to train a random forest model for generalized Supreme Court predictions [6]. These two studies boast an accuracy rate between 70 and 75% but interestingly, a hobbyist from Queens, New York and a legal correspondent at New York Times and Yale Law School claim to accuracy rates just as high without the use of a statistical model, instead relying on their memories of justices' past decisions, opinions, and oral arguments [7]. Jacob Berlove, the hobbyist from Queens, has correctly predicted Supreme Court cases for the past 3 years, more than 80% of the time.

The grounds for predicting Supreme Court Decisions are still evolving and current statistical models may not be capturing the full potential of predictability by relying solely on variables that arise out of the Supreme Court such as previous voting history and justice demographics and ideologies. The following is a non-exhaustive collection of factors that can be correlated with predicting Supreme Court decisions, as identified from existing research. The objective of the research presented in this paper is to expand upon current statistical models in order to incorporate more factors and analysis into predicting the decisions of the Supreme Court.

Predictor(s)	Corresponding Study
Justice and Court Background Information, Supreme Court Trends, Case Information	"Predicting the Behavior of the Supreme Court of the United States: A General Approach" Katz, Bommarito II, Blackman [6]
General Case Characteristics, Predictions from Legal Specialists	"The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking" Ruger, Kim, Martin, Quinn [5]
Justice Ideologies	"Not the Whole Story: The Impact of Justices' Values on Supreme Court Decision Making" Songer, Lindquist [4]
Case Information &, Precedent, Political Atmosphere, Oral Arguments	"Predicting Supreme Court Decisions" Rajendiran, Monica

Fig 1. Predictors of Supreme Court Decisions in existing Research

Normative Analysis

Outside of law, machine learning techniques that are able to predict and offer data-driven analyses are transforming industries such as finance and medicine. Within the law space, these techniques specifically applied to predicting Supreme Court decisions should have implications for how individuals, companies, lawmakers, and the government interpret how the U.S. judicial system upholds, maintains, and overturns laws and rules within the national legal framework. They should also change the landscape of how law is practiced, for the better.

The field of law is heavily influenced by individual decisions, regardless of whether the decision is in the role of judge, prosecutor, defender, or legislator. There are inherent biases in many, if not all, of the decisions regardless of how non-partisan or neutral the role demands the individual making the decisions to be. Successfully predicting court decisions based on factual information prior to the decision of the court should highlight these human biases and how the decisions of the courts are made. For example, Songer and Lindquist [4] found that in many cases, precedence influences the decisions of Supreme Court decisions, more so than personal ideology. And in other cases, the decisions of the Supreme Court justices are based on public opinion. A predictive model should highlight the sources of influences to the Supreme Court. Justices, attorneys, and lawmakers can be more aware of these influences when making decisions and be better informed in the decision making process.

Successfully incorporating machine learning techniques that can help predictions and interpretations of the law should also make the law more accessible to a wider audience. The law has a reputation for being a complex hierarchical labyrinth with multiple courts, procedures, regulations, jurisdictions, and other sometimes confusing terminology used to represent the proceedings between a prosecutor and defendant. It is important for the general population to be be involved in legal proceedings and the precedents for laws set by the courts, especially cases that significantly change the social, environmental, political, and financial landscape. Public involvement can incentivize lawmakers and judges to create laws that are more closely aligned with public opinion. For example, the website "FantasySCOTUS" has been heralded by Supreme Court Justice Stephen G. Breyer for creating public interest in the courts, saying "The more the public knows about the court, the better" [8]. FantasySCOTUS invites anyone to make predictions about cases held before the Supreme Court, stimulating interest in the legal process and judicial system which in turn leads to the creation of more informed public citizens that can generate input into how the legal landscape around important issues are shaped.

Hypothetically, predictive modeling of Supreme Court decisions could present a reality where attorneys and judges are replaced by an autonomous system of decision making. This might be far off in the future since such an autonomous system would need to be free of human bias in its creation, but the ability for a system to create rulings in a non-partisan, ideologically free way that is solely based on the legal framework of the Constitution is an imaginable scenario. However, such an autonomous ruling system would face much opposition by those whose jobs would be eliminated by its creation.

Stakeholders

The decisions of the Supreme Court have generally been regarded as fairly uncertain due to the amount of human input that is required, but the decision making process is still very closely followed not only by legal analysts and lawyers, but also by those interested in how the decisions of the courts affect the entire U.S. justice system. The ability to predict Supreme Court decisions has implications for many groups and individuals such as non-profit groups, government agencies, law schools, attorneys, and lawmakers.

For attorneys and individuals who find themselves tangled in the court system, a data driven approach to predicting court cases can greatly aid in guiding legal decisions. Legal decisions can be better informed, making the court system more efficient, transparent, and less expensive in terms of reducing legal fees. Along with individuals, companies can also better adjust their internal or external facing strategies in response to predicted legal decisions. However, there is a negative side to the ability to predict court decisions in that the predictions can create an unequal advantage favoring those with a greater amount of wealth or influence in legal disputes. Predictive tools should be open source, transparent, and easily accessible to the public.

Government agencies and nonprofit organizations can benefit from legal predictive tools in similar ways to individuals and attorneys. The biggest benefit to these groups is increased data and knowledge on which direction a court case may sway so that agencies can strategically prepare for the consequences of the outcome. The mere predictions of how a court may rule on a particular legal dispute may also affect initiatives or agendas for these agencies.

Law schools and their respective students can benefit from successful predictive methods of the Supreme Court in that they can better analyze what factors influence the decisions of the courts, including the biases and ideologies that come into the decision making process. This can better help them prepare for how they might litigate or analyze court cases in their professional careers. Academia in general can benefit from predictive models in court decision making in that research into the legal and judicial framework can be furthered through the quantification of what has been always seen as very human, complex, and ideological rulings. Research that looks at psychology in law, how justice ideologies change over time, and how rulings are affected by external and internal factors can also benefit from a model.

Legislators and politicians benefit from a predictive court decision model in that they can use the information and potential rulings to make time saving, efficient, and strategic decisions about which legislation to pursue or try to advance.

Impact

The envisioned impact of a predictive model for Supreme Court decision making is as follows:

Stakeholder	Impact
Individuals (General Public)	● Increased transparency into potential legal rulings ● Increased understanding of consequences of legal rulings ● Increased possibility of wealthier/influential individuals having an unfair advantage in legal proceedings
Attorneys	● Better guided and informed legal strategy (e.g. better litigation or settlement strategies)
Government Agencies/Non-profit Organizations	● Increased insights into how legal proceedings might affect themselves ● Increased insights into which strategies or initiatives to take based on legal proceedings
Law Students/Academia	● Advancement of research into quantifying or automating aspects of the judicial system ● Increased understanding of factors that influence the judicial system
Legislators/Politicians	● Better informed strategic decisions ● Efficiency and insights into which legislation to pursue or try to advance

Fig 2. Impacts to Stakeholders of Supreme Court Decision Making

2. Objectives and Metrics

Objectives

Factors I am trying to optimize, increase, decrease, or otherwise change in order to transition from the current state (as described in the descriptive analysis) to the alternative and desired situation (as described in the normative analysis), with regards to organizations and individuals that can benefit from predicting Supreme Court decisions are listed as following:

- Identification of factors that influence Supreme Court decision making

- Quantification of the influence that certain factors have on Supreme Court decision making
- Understand how predictable Supreme Court decision rulings are
- Increase accuracy and predictive power of current statistical methods that try to predict Supreme Court decision rulings
- Understand of how the judicial system works
- Incentivize public involvement into judicial system through increased transparency

Metrics

I would measure the following in order to satisfy our objectives and determine if my model successfully predicts Supreme Court decisions:

- Ability to predict whether or not a case affirms or reverses the decision of the lower court
    - Validated by error rate of prediction model when applied to validation set
        - False Positive Rate/True Positive Rate (ROC/AUC)
        - Confusion Matrices
    - Use of different methodologies to test the predictability of supreme court decisions (e.g. random forests, support vector machines, logistic regression)

3. Related Work

In order to understand the current state of how the decisions of the Supreme Court can be predicted, I looked at two studies that motivate my research into understanding the Supreme Court and its decision making process. In the following sections I describe these two studies in more detail to illustrate their relatedness to the work I am trying to do in this research.

3.1 Predicting the Behavior of the Supreme Court of the United States: A General Approach

In their study "Predicting the Behavior of the Supreme Court of the United States: A General Approach," Katz et al. created a methodology that has accurately predicted well over 70% of the Supreme Court's overall decision making from 1946 and onwards [6]. To many, this is considered the first study of its kind that merges computation and machine learning with the law. While there has been a strong interest in predicting Supreme Court decisions, notably for legal analysts, correspondents, and hobbyists, Katz and his team of researchers have opened up the field of Supreme Court predicting to analysts and data scientists.

The researchers designed a model in order to predict the voting behavior of the Supreme Court using randomized trees. They use more than 60 years of decisions made by the Supreme Court in order to correctly identify about 70% of the Supreme Court's decisions and over 70% of individual justice votes. The researchers find that their predictions are comparable to those of legal experts and scholars. What is unique in this model is that it is generalizable. Often, legal scholars or analysts depend on many circumstances surrounding a Supreme Court decision such as oral arguments or the current circumstances around the case in question. Katz and his team have created a method of using prior data that acts as precedent in order to determine whether the Supreme Court will affirm or reverse a court decision.

The researchers start their research off with the goal that they intend their model to be general, robust, and fully predictive. They intend on using random forests due to their problem being a classification task. They forecast whether the Supreme Court as a whole will decide to either affirm or reverse the judgement of a lower court, while also forecasting whether an individual justice will do the same. They rely on data from the Supreme Court database (SCDB) and build upon the data. The SCDB provides both justice and case information for each individual case. The researchers feature engineer many features such as information on the Current Supreme Court Trends (e.g. Mean Agreement Level of Current Court, Mean Current Court Direction), Overall Historic Supreme Court Trends (e.g. Mean Court Direction Circuit Origin, Mean Court Direction), Lower Court Trends, Individual Supreme Court Justice Trends, and Differences in Trends.

Ultimately, the researchers used 7,700 cases and 68,000 individual justice votes over the course of 60 years. The researchers found that their predictive model correctly forecasted 69.7% of the Case Outcomes and 70.9% of Justice Level Vote Outcomes over the sixty year period. This study is considered a seminal study in the field of computational law, especially with relation to the Supreme Court, so the work that I perform as described below will be very influenced by their work. However, I intend to deviate through separate feature engineering of my own.

3.2 The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking

This 2014 study focused on using an interdisciplinary approach to compare political science and legal approaches to forecasting Supreme Court decisions [5]. The researchers obtained predictions for every argued case during the 2002 term, prior to oral arguments, using a statistical model that used general case characteristics and another model that used a set of independent predictions by legal analysts or scholars. The researchers found that the statistical model performed better than the legal experts in forecasting the outcomes of the 2002 term decisions.

The statistical model predicted 75% of the Court's affirm/reverse results correctly, while the legal analysts and scholars collectively got near 60% of the court outcomes correctly. The researchers noted that the results of the statistical model were notable because unlike the information used by the legal scholars and analysts, the statistical model did not use information about the specific laws or facts that surrounded the case. The success of the model primarily came from being able to predict the votes of more ideologically extreme Justices. The model was found to have more difficulty in predicting the centrist Justices. Notably, the model also had varying degrees of success when it came to predicting cases by issue. The model was good at forecasting economic activity issues while the legal scholars and experts did better at predicting cases that were related to "judicial power" issues.

The statistical model of the researchers depended on six variables: circuit (court) of origin, issue area of the case, type of petitioner, type of respondent, ideological direction of the lower court ruling, and whether the petitioner argued that a law or practice was unconstitutional. The variables allowed the model to generate a predicted vote for each Justice and a predicted outcome for each case before it was argued in the 2002 term.

4. Define Hypotheses and Approach

I explore how to predict Supreme Court decisions using existing case information obtained from the Supreme Court Database (SCDB). I use many of the variables that existing Supreme Court predictors use, with the addition of analysis and inclusion of variables from the political environment, (e.g. the makeup of Congress) or variables indicating economic health (e.g. unemployment rate). I supplement this with data exploration that aims to identify any trends, patterns, or interesting findings from the data. I intend on creating a generalizable prediction model so my evaluation metrics will be tested on chronologically different Supreme Court terms and cases. I am interested in this problem because I have a strong interest in the laws that we have created, how they are enforced and interpreted, and how they impact the country, especially those laws that pertain to social or environmental issues. I see the ability to successfully automatically predict court cases as one that can lead to a changed interpretation of how legal experts, scholars, and justices themselves see the law and the effect that human biases have on it.

Hypotheses

Hypothesis A:

I hypothesize that the data obtained from the Supreme Court Database about general court case information can reveal trends, patterns, or interesting findings about the Supreme Court decision making process.

H_0: The case data obtained from the Supreme Court Database cannot reveal trends, patterns, or interesting findings about the Supreme Court decision making process.

H_A: The case data obtained from the Supreme Court Database can reveal trends, patterns, or interesting findings about students, specifically clear correlations.

Hypothesis B:

I hypothesize that the data obtained from the Supreme Court Database about general court case information, along with featured engineered variables, can predict the decision of the Supreme Court as a whole on a case (e.g. whether the Court affirms or reverses the lower court judgement).

H_0: The historical case data provided by the Supreme Court Database cannot predict the decision of the Supreme Court on a case.

H_A: The historical case data provided by the Supreme Court Database cannot predict the decision of the Supreme Court on a case.

Hypothesis B is predictive in that I subset particular variables from the Supreme Court case data to determine whether whether the Supreme Court will affirm or reverse the judgement of the lower court. Hypothesis A is of an exploratory nature due to the data being technical in terms of legal terminology and procedure, as well as offering a chance to explore aspects of the decision making process and nature of the Supreme Court. Based on previous literature in this field, this study differentiates itself in that it is using historical case information along with engineered features extracted from economic, temporal, or political factors.

Data

This study depends on data obtained from the Supreme Court Database [9]. The SCDB is hosted on Washington University Law School's website. It has data ranging from 1946 to 2015, over six decades of case and justice data. The data is coded by legal experts and scholars and is regularly maintained and routinely subjected to reliability analyses by a myriad of academic groups and studies. For each case, there exists up to 247 variables including background variables, chronological variables, substantive case information, outcome variables, voting variables, and opinion variables (relating to the Justices).

Although both justice and case background information exists for each Supreme Court case, this study will be focusing on the case background information. This is for several reasons including: the primary statistical models that have predicted Supreme Court decisions have been built on this data and I want to see if I can expand upon it successfully, the data contains almost 250 features so I want to focus on successful feature reduction and feature engineering, and I am more interested in creating a generalizable and robust model using case information over justice information.

Case information includes information such as the *issue*, *lawType*, and *caseSource* which can impact how justices may vote. The information can also be used to engineer features such as *monthArgument* which keeps track of what month an oral argument was argued before the courts. The original case data set contained close to 250 variables. It was pared down to variables that were determined through a combination of information contained in the literature review and what I was most interested in seeing connected to predicting Court decisions. For example, I removed all "outcome" variables (e.g. *decisionDirection* apart from the variable that directly determined the decision of the court (*caseDisposition*), since including those variables in the data would be similar to keeping *caseDisposition* in the data set. A significant amount of time was spent analyzing the individual variables representing a particular case and making sure they were suitable for inclusion in the predictive model. Cases were only selected where judges had given an opinion (so no per curiam cases or cases with opinions written by others were considered). Only cases that judges had explicitly affirmed or reversed the judgement of were considered. There were also a significant amount of null values in the data which required managing them in a way such that the sample size of case data would be sufficient for modeling and testing. A total of 13,266 cases were originally made available but after paring down to account for inconsistencies or null values in the data, 5,842 observations of cases were left with 30 variables (including feature engineered ones). The following is a list of the predictors that were used for the creation of the predictive statistical model for determining whether the Supreme Court affirmed or reversed the judgement of the lower court. Engineered features are listed in the "Predictor" column with an indication of "[Engineered]".

Predictor	Type	Description
dateDecision	Date	Date when Court Decision was made
term	numeric	Supreme Court Year (October - June)
naturalCourt	numeric	Period during which no personnel change occurs; values divide the Courts into strong natural courts, each of which begins when the Reports first specify that the new justice is present but not necessarily participating in the reported case. Similarly, a natural court ends on the date when the Reports state that an incumbent justice has died, retired, or resigned

chief	factor	Chief Justice of Supreme Court
dateArgument	Date	day, month, and year that the case was orally argued before the Court
jurisdiction	factor	The Court uses a variety of means whereby it undertakes to consider cases that it has been petitioned to review. The most important ones are the writ of certiorari, the writ of appeal, and for legacy cases the writ of error, appeal, and certification.
lcDisagreement	factor	Indicates that the Supreme Court's majority opinion mentioned that one or more of the members of the court whose decision the Supreme Court reviewed dissented
certReason	factor	Provides the reason, if any, that the Court gives for granting the petition for certiorari
lcDisposition	factor	Whether the court below the Supreme Court---typically a federal court of appeals or a state supreme court---affirmed, reversed, remanded, etc. the decision of the court it reviewed---typically a trial court
lcDispositionDirection	factor	Specifies whether the decision of the court whose decision the Supreme Court reviewed was itself liberal or conservative as these terms are defined in the direction of decision variable (decisionDirection)
precedentAlteration	factor	A "1" will appear in this variable if the majority opinion effectively says that the decision in this case "overruled" one or more of the Court's own precedents
voteUnclear	factor	Votes in a case are those specified in the opinions
issueArea	factor	Separates the issues identified in the preceding variable (issue) into the following larger categories (14 categories)
authorityDecision1	factor	Specify the bases on which the Supreme Court rested its decision with regard to each legal

		provision that the Court considered in the case (see variable lawType)
splitVote	factor	Indicates whether the vote variables (e.g., majVotes, minVotes) pertain to the vote on the first or second issue (or legal provision). Because split votes are so rare over 99 percent of the votes are on the first issue
majVotes	numeric	Specifies the number of justices voting in the majority
minVotes	numeric	Specifies the number of justices voting in the dissent
vote	factor	Identifies the issue for each decision
opinion	factor	Indicates the opinion, if any, that the justice wrote
direction	factor	Indicates whether the justice cast a liberal (2) or conservative vote (1)
majority	factor	If vote is in dissent (1) or majority (2)
senate_dems [Engineered]	numeric	Number of senate democrats in term
senate_reps [Engineered]	numeric	Number of senate republicans in term
house_dems [Engineered]	numeric	Number of house democrats in term
house_reps [Engineered]	numeric	Number of house republicans in term
unemployment_rate [Engineered]	numeric	Unemployment rate for term
monthArgument [Engineered]	numeric	Month oral argument was received for case
monthDecision [Engineered]	numeric	Month decision was made for case
dateSpan [Engineered]	numeric	Days between oral argument and decision
Response	**Type**	**Description**
caseDisposition [Engineered]	binary	1 - Affirms; 0 - Reverses

Fig 3. Predictors and Response Variables for Models

Biases

There are many potential sources of bias for this study. The most glaring of these is that the majority of data feature selection, wrangling, and cleaning is based upon my knowledge of law and what is important to the decisions of the Supreme Court. I do not have the expertise of a legal scholar or analyst and only depended on my research to assess what was crucial to be included in the data. I also chose variables that had relative minimum sparsity and a number of missing values below 3000 in the overall data table. I also cleared a lot of variables (from a little over 200 down to 300 including featured engineers). My predictions may be biased in that I could not include all the variables from the Supreme Court Database due to factors such as computational feasibility, previous research indicating certain variables were more indicative of predictive success over others (*lcDisposition*), high sparsity (e.g. *issue*), and irrelevancy (e.g. *naturalCourt*).

5. Hypotheses, Methods, and Evaluation

Hypothesis A (Exploration):

H_0: The case data obtained from the Supreme Court Database cannot reveal trends, patterns, or interesting findings about the Supreme Court decision making process.

H_A: The case data obtained from the Supreme Court Database can reveal trends, patterns, or interesting findings about students, specifically clear correlations.

Methods & Assumptions

I began my analysis by examining the missingness contained in the Supreme Court case information, to give me some sense of which features I will not be able to use in my analysis. I aimed to keep the most intact data set while preserving all the essential features. I started off with 13,266 observations and after selecting the variables that I wanted to focus on (based on literature review and what I believed would be interesting to connect with predicting Supreme Court decisions), I then filtered for only Supreme Court cases that had actually been voted on. For example, there are a number of cases where the Supreme Court does not actually affirm or reverse the judgement of a lower court. There are times where they offer suggestions and send the case back to the lower court or have someone else write the opinion without a vote. I only wanted to select cases that the Supreme Court had actually voted on. In total, that narrowed the number of cases I could work with to 5,842. After that I spent time feature engineering and looking at interesting trends or features of the data, as listed below.

Missingness

Missingness, especially, involving outcome variables to a Supreme Court decision make working with the data difficult. Increased sparsity can also make the outcomes of your predictions and interpretations within the statistical model not as significant or powerful. The following graph depicts where missingness exists within the Case dataframe (as pulled from the Supreme Court Database). The missing values are in light gray.

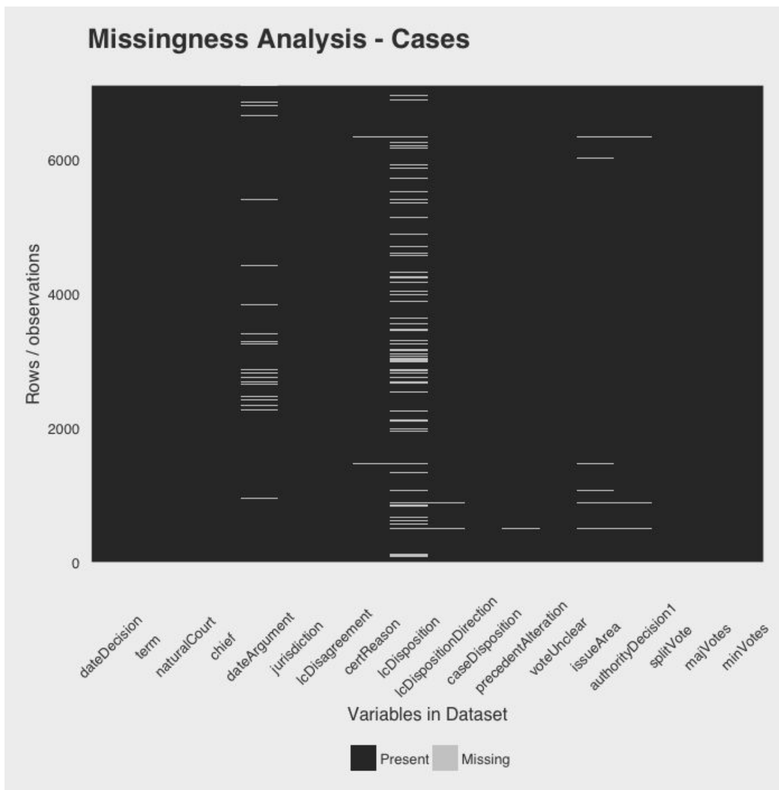


Fig 4. Analysis of Missingness in Case Information

Cases Per Term

A term is defined as the year that the Supreme Court hears and decides on cases. This 'year' spans from October to June. The following graph depicts the number of cases that are heard per term. Based on the slight differentiation of column spacing, we can see that this is most likely a result of missing data for some of the term years. Notably, this graph is in line with research that says that since the 1980s, the Supreme Court has taken on less and less cases.

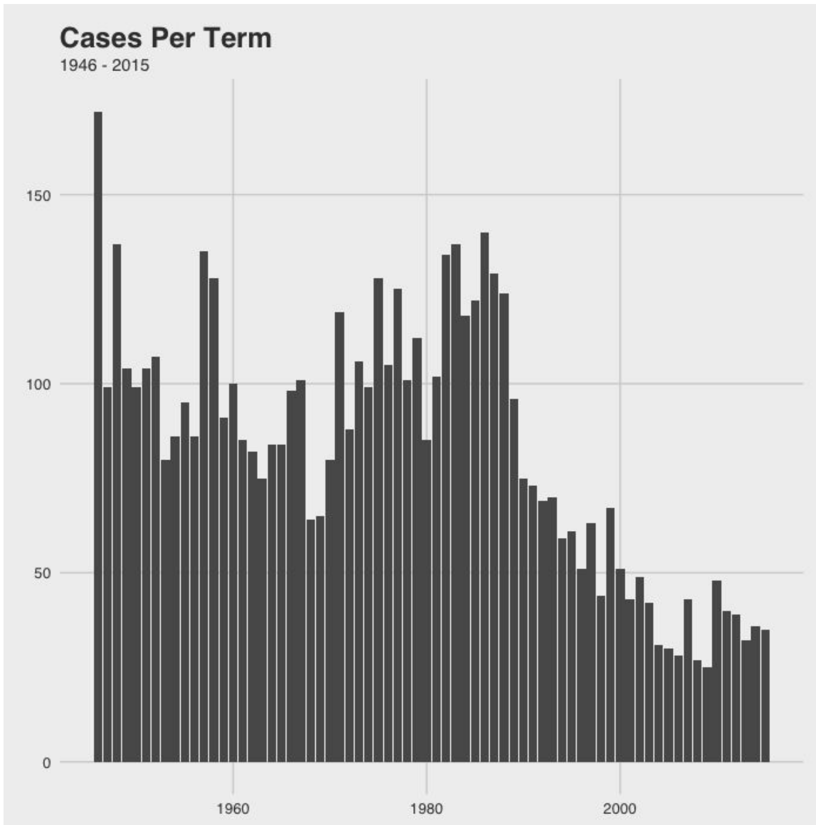


Fig 5. Cases per Term (1946 - 2015)

Average Days Between Argument and Decision

The Supreme Court hears oral arguments where both the petitioner and respondent in the case have a chance to explain their positions. While they are defending their positions, the Supreme Court justices also have a chance to ask questions and clarify any questions. These questions are often windows into the decision making process and offer insights into how certain justices may be inclined to vote. After the oral arguments, the Supreme Court justices then meet to make a decision. This span, labeled as engineered feature *dateSpan* in the data is averaged for each term to produce the following graphs:

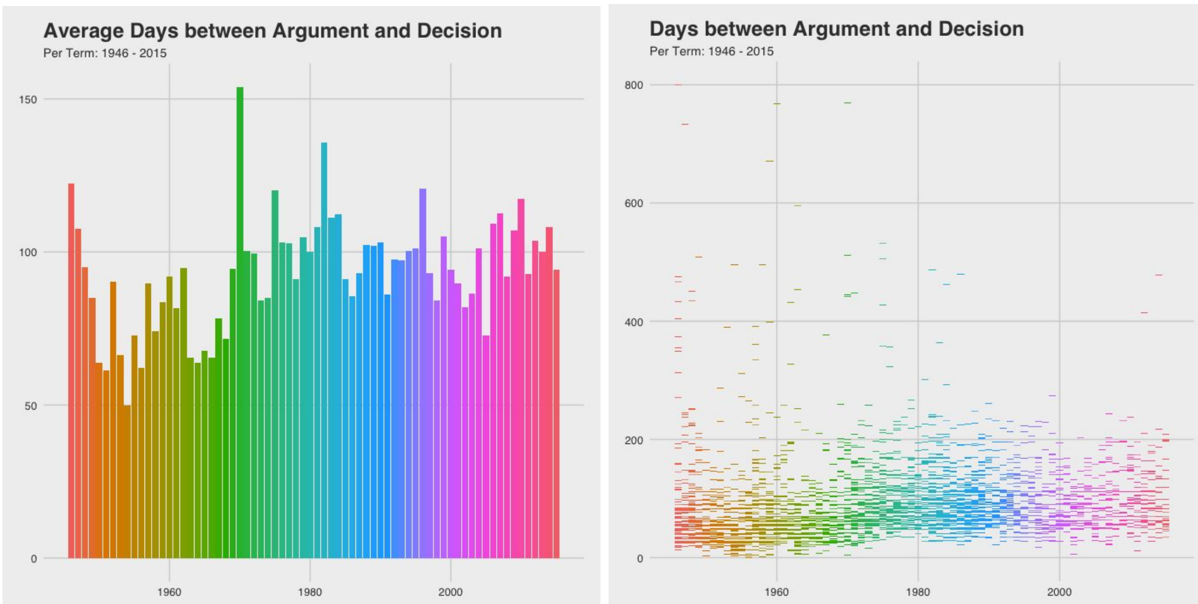
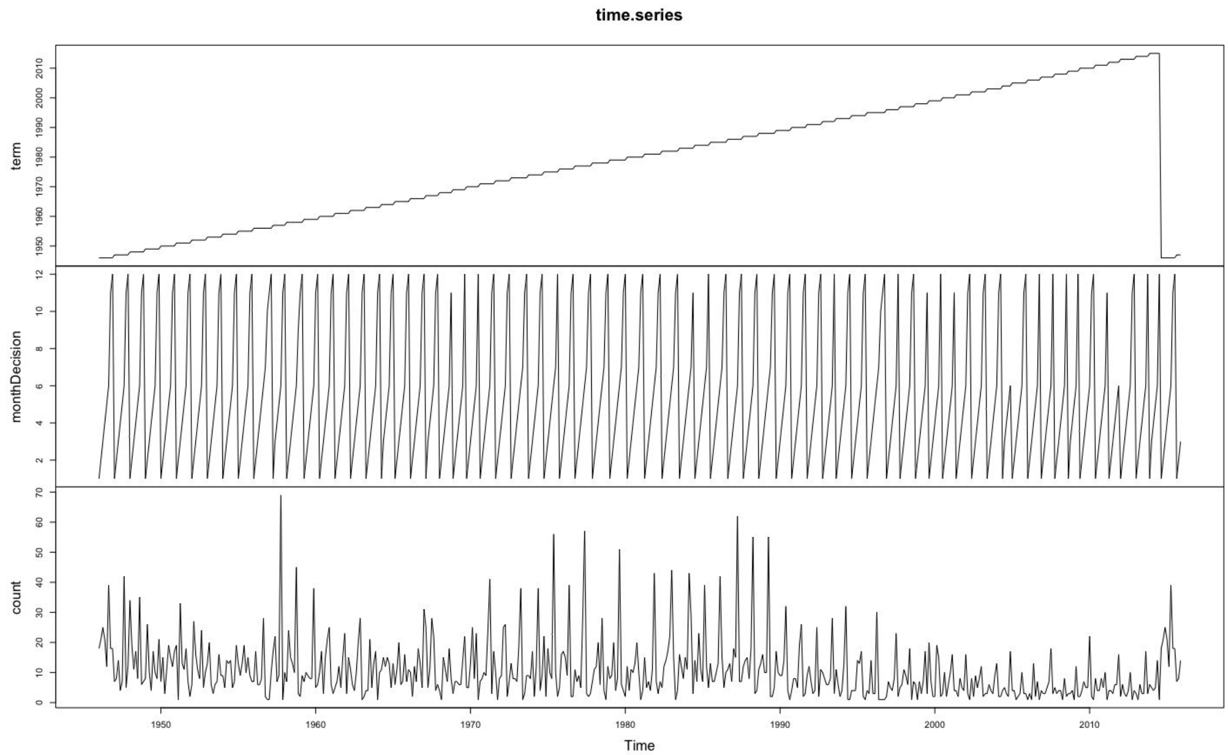


Fig 6. Two different visuals of the Days between an Argument and Decision (Avg on Left)

Time Analysis

As part of the data exploratory part for this data, I briefly examined a part of the case information that could be interpreted as a time series object. The following shows the months that a decision was made along with the term that the month falls in. The Supreme Court falls into 8 month cycles (October to June) so the second time series graph shows these cycles as occurring every 8 months. There does not appear to be a cyclic pattern for the graph that shows the count and the term.



Text Analysis

As mentioned above, the Supreme Court hears oral arguments where both the petitioner and respondent in the case have a chance to explain their positions. While they are defending their positions, the Supreme Court justices also have a chance to ask questions and clarify any questions. These questions are often windows into the decision making process and offer insights into how certain justices may be inclined to vote. Transcripts are available for Supreme Court oral arguments (although they are often behind a paywall). The following is a simple text analysis of the text found in oral argument transcripts between 2001 and 2002.

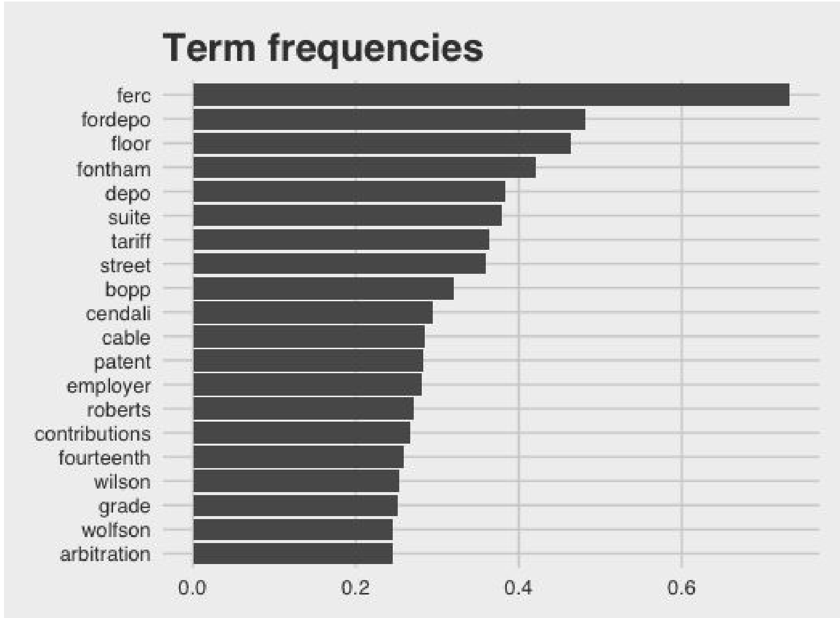


Fig 7. Top 20 Most Frequently Occurring Words in 2001 - 2002 Oral Arguments

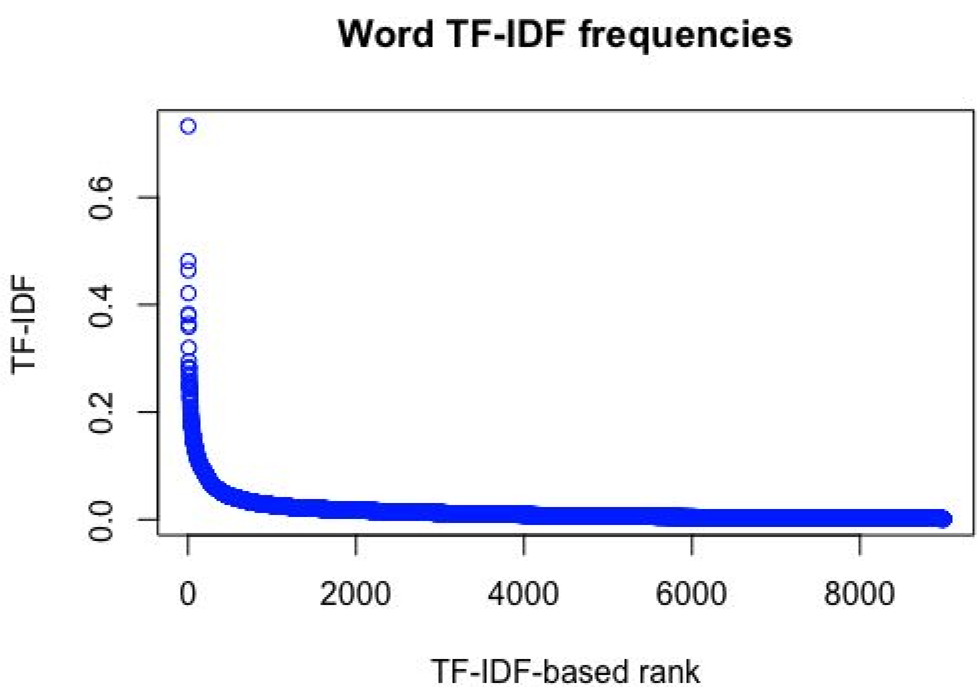


Fig 8. Plot of Ordered Words by Frequency

Hypothesis A Evaluation

Based on my exploratory analysis of the missingness, distributions, and variations within a range of variables, I have reason to believe that it should be possible to build a model which gives a satisfactory portrayal of the Supreme Court making process, if not one that can predict whether the decisions of the court. There is a clear downward trend in the number of cases that the Supreme Court has taken on since the 1980s, as research suggests. It will be interesting to see if this somehow affects the predictability of decisions. There does not appear to be high variation in the span of days between the date of the oral argument and the date of the Supreme Court decision when looking at the averages. However, the outliers or values that deviate from the average may affect the predictability of decisions. It does not appear that the text analysis will help in predicting Supreme Court decisions because transcripts are only available for the terms 2001 - 2002. However, it was still an interesting analysis to see what the top words showing up in the transcripts were.

Hypothesis B (Binary Classification):

H_0: The historical case data provided by the Supreme Court Database cannot predict the decision of the Supreme Court on a case.

H_A: The historical case data provided by the Supreme Court Database cannot predict the decision of the Supreme Court on a case.

Methods & Assumptions

In this paper I attempt to predict the binary response of whether the Supreme Court would affirm or reverse the judgement of the lower court through a variety of methodologies.

A total of 13,266 cases were originally made available but after paring down to account for only cases that the justices had given an opinion on and null values in the data, 5,842 observations of cases were left with 30 variables (including feature engineered ones). The featured engineered data included the unemployment rate per term, number of republicans and democrats in the house and senate, and the number of days between an oral argument and the date of decision over a case.

I was interested in determining which model, out of a variety of models, would perform the best in predicting the Supreme Court decisions. The methodologies I used were logistic regression, random forests, and support vector machines.

Logistic Regression

Logistic regression models do not follow all of the more well known assumptions of linear regression models, such as the assumption of linearity between the explanatory variables and the response variable. Some assumptions for logistic regression models are as follows:

1. Linear relationship between the logit of the explanatory variables and the dependent variables
    a. I assume that the logit of our explanatory variables and our response variables have a linear relationship
2. A large sample size (reliability of estimation significantly declines with a small sample)
    a. I took steps to ensure that we could get the maximum sample size for our data. After accounting for variables that were irrelevant for the purposes of our study and the missingness in the data, the sample size was 5,842.
3. Multicollinearity does not exist
    a. I checked for multicollinearity using the VIF function.
4. The independent variables do not need to be an interval level
    a. I have independent variables that are numeric such as number of majority or dissenting votes are measured.

Logistic regression is useful for classification. It can also model the probability that the response belongs to a particular category. After creating the logistic regression model with all of the variables that we feature selected and engineered, I used hybrid stepwise selection (forward and backward) to select for the features. The resulting model showed that all variables were significant so I did not remove any (to reduce the possibility of overfitting). The cross-validated (20-fold) prediction accuracy of this method was 68.4%.

To further investigate our logistic regression model, I produced an ROC curve. I held out 25% of the observations at random as a test set in order to evaluate testing performance. The ROC curve for these predictions can be found in the figure below.

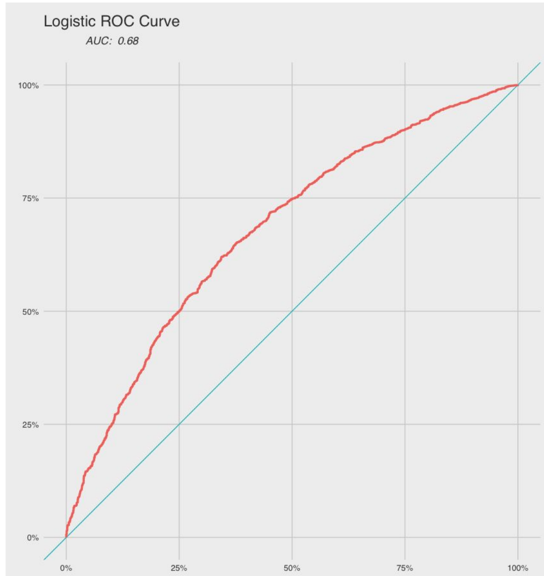


Fig. 9 ROC Curve for Logistic Regression Model

As the Figure demonstrates, the area under the ROC Curve (AUC) was 0.68. This indicates that our model improves upon random guessing, although marginally.

Random Forests

Random Forests offer an improvement over decision trees and bagged trees by decorrelating the trees. Whenever a split in the tree is performed, a random sample of predictors are chosen in the consideration of the split. I started out with creating randomized training and testing sets so that the case data was split 75% and 25% respectively. I also separately created a random forest model that was trained on not randomly splitting the data 75% and 25% and instead approximating the same ratio but using chronological data (e.g. the 25% was cases after 1990 while the 75% was about cases that were before 1990). The second random forest with the chronologically separated training and testing data performed much worse than the random forest that was created using randomly split data.

Confusion Matrix (Random Forest 1)		
	Reference	
Prediction	0 (Reverse)	1 (Affirm)
0	363	72

1	228	798

Fig. 10 Confusion Matrix for Random Forest on Randomly Split 75/25 Training/Testing Data

Confusion Matrix (Random Forest 2)		
	Reference	
Prediction	0 (Reverse)	1 (Affirm)
0	363	72
1	228	798

Fig. 11 Confusion Matrix for Random Forest on Chronologically Split 75/25 Training/Testing Data

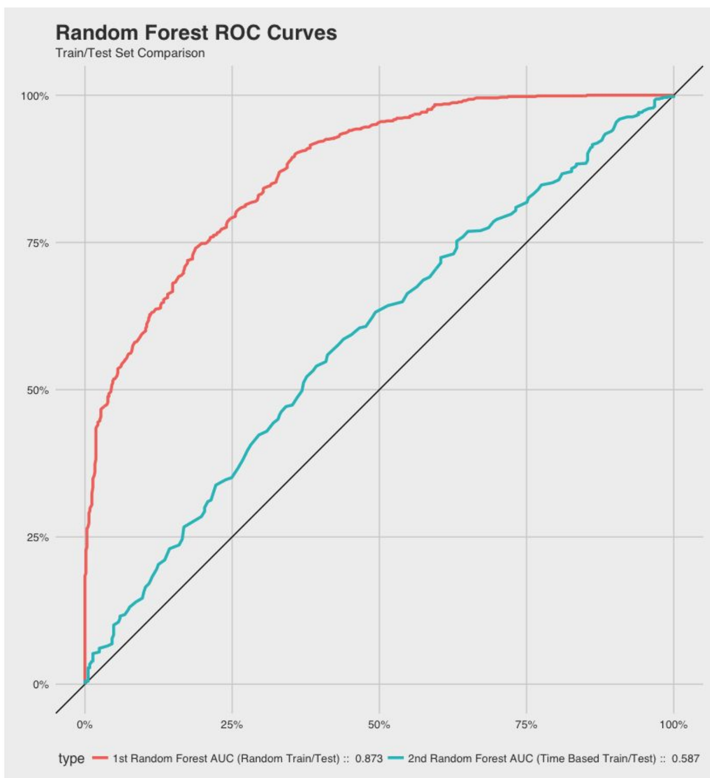


Fig. 12 Comparing ROC curves/AUC scores for the different Random Forests trained on different splits of testing/training data

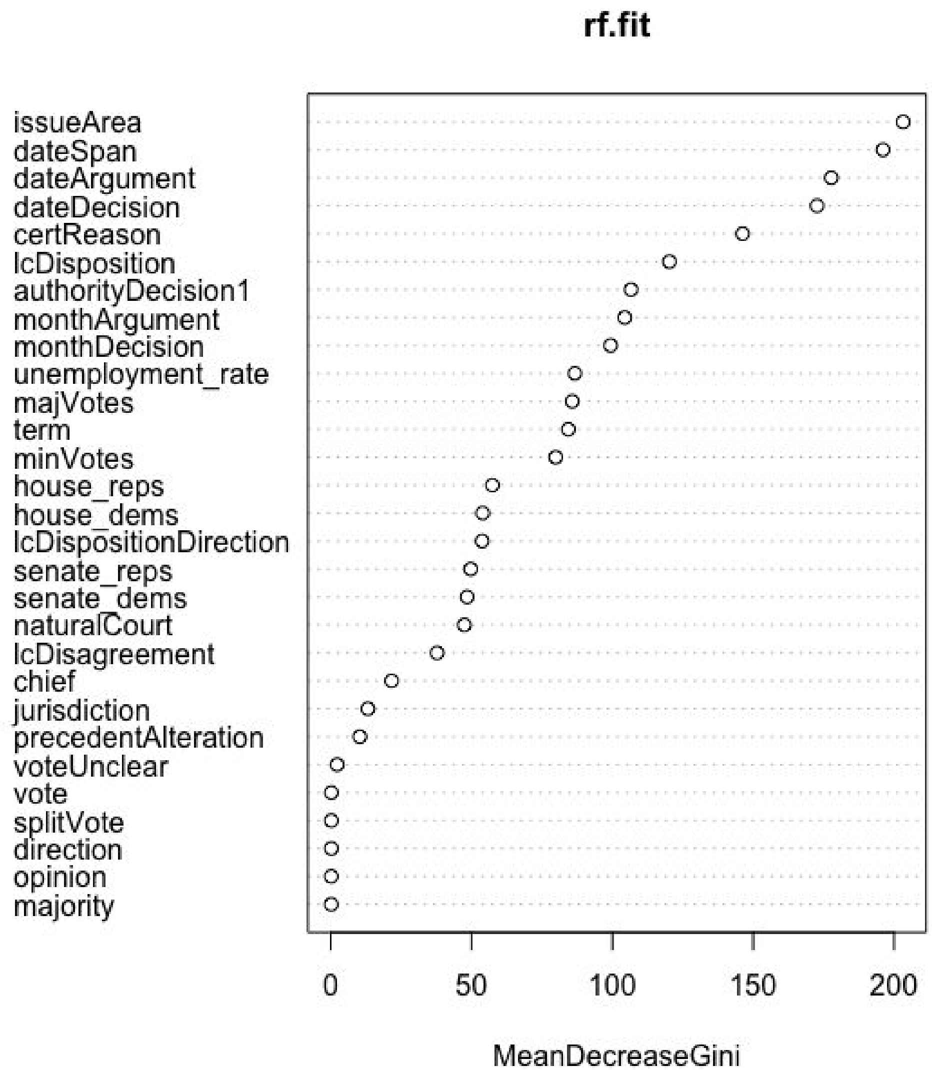


Fig. 13 Importance Plot for Random Forest (#1)

From looking at the Importance Plot we can see that the most useful variables are issueArea, dateSpan, and Date Argument. Our raw accuracy for our well performing Random Forest (trained on randomly split 75/25 data) shows that our raw accuracy is 79.8% accuracy. From looking at our confusion matrix it does not appear that our data is heavily skewed towards either case disposition but our Sensitivity is 0.3406 while our Specificity is 0.7030. We would like our Sensitivity (True Positive Rate) to be higher.

Support Vector Machines

Support Vector Machines are discriminative classifiers that are defined by having a separating hyperplane between whatever data is needed to be separated. I started out with creating randomized training and testing sets so that the case data was split 75% and 25% respectively. The raw prediction accuracy for this model was 64.75%, not performing as well as the better performing random forest model.

Hypothesis B Evaluation

The methods used to evaluate Hypothesis B (predicting whether a case would be affirmed or reversed by the Supreme Court) were logistic regression, random forest models, and support vector machines. The predictors for these models are provided on page 11, with the response variable being a binary indicator of whether the Supreme Court affirmed or reversed the judgement of the lower court. These methods were evaluated using 20 fold cross fold validation, variable importance plots, confusion matrices, raw prediction accuracy, and AUC scores. Random Forest performed the best out of Logistic Regression and Support Vector Machines. The Random Forest had a score of .87, the Logistic Regression model had a score of .67, and the SVM had a score of .66. Comparing the Random Forest and SVM, Random Forest had a raw accuracy score of 79.8% while SVM had a raw accuracy score of 64.8%. According to our confusion matrix, our sensitivity is satisfactory (0.6639) so these results do support our hypothesis that we can predict whether a Supreme Court rules to affirm or reverse a decision based on historical case information.

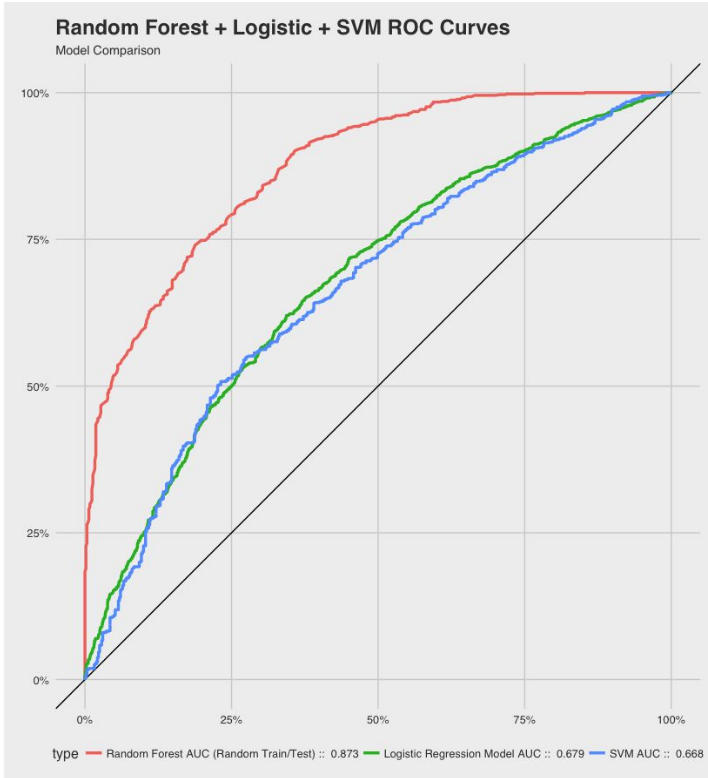


Fig 14. Comparison of Random Forest, Logistic Regression, and SVM AUC scores

6. Conclusions and Results

Data

Data from the Supreme Court Database was investigated in an effort to understand what can be inferred about the Supreme Court decision making process and how decisions are made. Specifically, data was gathered from the historical case information for each case. There were close to 250 variables for each case. Many of these variables had to be pared down in order to select the ones that, based on literature review and personal bias, would be the most adequate for predicting Supreme Court decisions. I am not a legal expert and performed variable selection based on my own review of existing literature on law and statistical models when predicting within law. This inherently leads to human biases. This highlights the importance of having a subject matter expert on hand when working with data that is in a very specific field, even though the field still lends itself to statistical and machine learning methods. This data also highlights how important it is to have "good" data to perform significant analyses that can produce worthy results. The Supreme Court Database is well relied upon for academic studies and as a result goes through rigorous review. For analysis on oral arguments, the data for argument transcripts was not so readily available. The official supreme court site hosting the arguments encouraged people to buy the transcripts so they were not even available for scraping and often copying. Significant insights could have been gained from the analysis on oral arguments if it were not behind a paywall.

Methods & Assumptions

I decided to see what could be extracted from a basic analysis of Supreme Court case information. I was hoping to extrapolate trends, patterns, or interesting information that could aid in the classification of the Supreme Court either affirming or reversing a decision. The exploratory data analysis gave me insights into features to engineer for Hypothesis B (binary classification issue) and the insights gained also helped give me a better idea of the Supreme Court decision making process and how that process has changed over time. Exploratory data analysis included use of text mining, in particular the TD-IDF matrices, and simple use of graphical and visual techniques to display information in a more digestible way.

The methods used for binary classification was logistic regression (using hybrid stepwise selection), random forest classifiers, and support vector machines. I was interested in comparing the results of different classifiers on the same data set in order to compare their performance. Categorical classifiers are unlike linear assumptions so the formal assumptions for logistic regression, random forests, and support vector machines have been minimally assessed so I approach this analysis with a degree of skepticism.

Evaluation

Based on my exploratory analysis of the missingness, distributions, and variations within a range of variables, I have reason to believe that it should be possible to build a model which gives a satisfactory portrayal of the Supreme Court making process, if not one that can predict whether the decisions of the court. There is a clear downward trend in the number of cases that the Supreme Court has taken on since the 1980s, as research suggests. It will be interesting to see if this somehow affects the predictability of decisions. There does not appear to be high variation in the span of days between the date of the oral argument and the date of the Supreme Court decision when looking at the averages. However, the outliers or values that deviate from the average may affect the predictability of decisions. It does not appear that the text analysis will help in predicting Supreme Court decisions because transcripts are only available for the terms

2001 - 2002. However, it was still an interesting analysis to see what the top words showing up in the transcripts were.

The methods used to evaluate Hypothesis B (predicting whether a case would be affirmed or reversed by the Supreme Court) were logistic regression, random forest models, and support vector machines. The predictors for these models are provided on page 11, with the response variable being a binary indicator of whether the Supreme Court affirmed or reversed the judgement of the lower court. These methods were evaluated using 20 fold cross fold validation, variable importance plots, confusion matrices, raw prediction accuracy, and AUC scores. Random Forest performed the best out of Logistic Regression and Support Vector Machines. The Random Forest had an AUC score of .87, the Logistic Regression model had a score of .67, and the SVM had a score of .66. Comparing the Random Forest and SVM, Random Forest had a raw accuracy score of 79.8% while SVM had a raw accuracy score of 64.8%. According to our confusion matrix, our sensitivity is satisfactory (0.6639) so these results do support our hypothesis that we can predict whether a Supreme Court rules to affirm or reverse a decision based on historical case information.

Results & Considerations

The results of this study do indicate that we can predict whether the Supreme Court will affirm or reverse the judgement of a lower court. Our preliminary data analysis showed insights that were then used in the building of predictive models using random forests, logistic regression, and support vector machines. Random Forests performed better than Logistic Regression and Support Vector Machines. As stated earlier, the Random Forest had an AUC score of .87, the Logistic Regression model had a score of .67, and the SVM had a score of .66. Comparing the Random Forest and SVM, Random Forest had a raw accuracy score of 79.8% while SVM had a raw accuracy score of 64.8%.

There are myriad ways to extend this project. This study was just scratching the amount of data that could be augmented to this study to predict Supreme Court decisions. For example, this study did not use information on individual justices but that information (consisting of things like ideological precedent) could vastly improve the current statistical models that attempt to predict Supreme Court decisions. Legal correspondents have said that they are able to predict (without statistical models) the decisions of the Court from hearing the oral arguments. The addition of transcripts for every Oral arguments could be a very predictive feature to add.

Steps Toward the Normative Scenario

The predictive models created in this study provide a foundation for further work in predicting Supreme Court decisions through historical case information and with new features such as economic or political indicators. This work can be further expanded by including information from oral arguments into the predictive modeling, as well as information on the individual justices themselves. The research presented in this paper offers additional support to existing statistical models and research that show that one can predict whether Supreme Court justices will affirm or reverse the judgement of the lower courts. However, a truly useful predictive model that can help individuals in the court systems, attorneys, law students, and companies looking to navigate their way through the legal system can only benefit from a predictive model that is free from human biases. As discussed in the introduction, the field of law is heavily influenced by individual decisions, regardless of whether the decision is in the role of judge, prosecutor, defender, or legislator. There are inherent biases in many, if not all, of the decisions regardless of how non-partisan or neutral the role demands the individual making the decisions to be. Successfully predicting court decisions based on factual information prior to the decision of the court should highlight these human biases and how the decisions of the courts are made.

7. References

1. Hamilton, Eric. "Politicizing the Supreme Court." Stanford Law Review Online 65 (2012): 35.
2. Ignagni, Joseph A. "Explaining and predicting Supreme Court decision making: The Burger Court's establishment clause decisions." *Journal of Church and State* 36.2 (1994): 301-327.
3. Casillas, Christopher J., Peter K. Enns, and Patrick C. Wohlfarth. "How public opinion constrains the US Supreme Court." *American Journal of Political Science* 55.1 (2011): 74-88.
4. Songer, Donald R., and Stefanie A. Lindquist. "Not the Whole Story: The Impact of Justices' Values on Supreme Court Decision Making." *American Journal of Political Science* 40.4 (1996): 1049-1063.
5. Ruger, Theodore W., et al. "The Supreme Court forecasting project: Legal and political science approaches to predicting Supreme Court decisionmaking." Columbia Law Review (2004): 1150-1210.
6. Katz, Daniel Martin, Michael James Bommarito, and Josh Blackman. "Predicting the behavior of the supreme court of the united states: A general approach." Available at SSRN 2463244 (2014).
7. Greenhouse, Linda. "Press room predictions." Perspectives on Politics 2.04 (2004): 781-784.
8. Fantasy Scotus https://fantasyscotus.lexpredict.com/about/
9. Washington University of Law - Supreme Court Database http://scdb.wustl.edu/data.php