Chapter 1

Models Versus Experts

Always listen to experts. They'll tell you what can't be done and why. Then do it.

– Robert Heinlein

Contents

On March 4, 1990, the *New York Times* announced that a Princeton University economics professor can predict the quality of wine without tasting a single drop. This professor believes that his predictions are more accurate than those of the world's most influential wine critic, Robert Parker, who called the predictions "ludicrous and absurd." These predictions have nothing to do with assessing the aroma, looking at the color, or determining the flavor profile of the wine; they are the results of a mathematical model.

Meanwhile, in a completely unrelated field, a study was being performed that ultimately reported shocking statistics about the quality of healthcare in the United States. One of the results of this study was the discovery that over half of 2,000 diabetic adults did not have a dilated eye exam in the past year, although diabetes is the leading cause of new cases of blindness among adults. They stated that "the dominant finding of our review is that there are large gaps between the care people should receive and the care they do receive." In research at the Massachusetts Institute of Technology, two analytics professors and a physician developed a mathematical model to assess the quality of care of patients in a fraction of the time that it takes a physician.

Mathematical models have even disrupted the fields of political science and law. In 2004, two law professors and two political science professors published an article claiming that a statistical model is better at predicting the results of Supreme Court cases than the collective opinions of experts. They predicted the affirm/reverse decisions of every Supreme Court case in the October 2002 term, cases including deep issues such as the constitutionality of affirmative action and various free speech rights. Lawyers, legal academics, and specialized journals that follow the Court closely have had little success in consistently predicting Supreme Court decisions in advance. However, these professors claim that the unbiased and unemotional nature of a model can better capture the decisions of the court.

In this chapter, we explore the possibility that predictive mathematical models can outperform expert human judgment by analyzing three different examples. While people are not always consistent in their opinions, are often emotional, and get tired, models are consistent, unemotional, and fast. We will show in this chapter that these characteristics lead to a competitive edge when models are used, especially if they are constructed using human expertise and judgment.

1.1 Predicting the Quality and Prices of Wine

Orley Ashenfelter made his prediction for red wine that is produced in the Bordeaux region of France, commonly referred to as "Bordeaux wine." He sought to address the mystery that while this wine has been produced in much the same way for hundreds of years, there are differences in price and quality from year to year that are sometimes very significant. Bordeaux wines are widely believed to taste better when they are older, so there is an incentive

to store young wines until they are mature. The main problem is that it is hard to determine the quality of the wine when it is so young just by tasting it since the taste will change so significantly by the time it will actually be consumed. This is why wine tasters and experts are helpful; they taste the wines and then predict which ones will be the best several years into the future. However, Ashenfelter observed that "bad" vintages are usually overpriced when they are young, and "good" vintages may sometimes be underpriced when they are young. Realizing that the advice of experts was making the market for young wines inefficient, he developed a different system for judging the wines.

Ashenfelter discovered two explanations for the variation in prices: the age of the vintage, and the weather. He hypothesized that since older wines have been held longer, they must be more expensive than younger wines. The fact that the quality of the grapes depends on the weather was widely understood. However, Ashenfelter pointed out that weather in the Bordeaux region can vary dramatically from one year to the next. Figure 1.1 is a scatterplot of the average growing season temperature (in degrees Celsius) versus the amount of harvest rain (in milliliters) in Bordeaux from 1952-1980. Each point is a year, and the years with higher than average prices are shown as triangles. (A year is marked as higher than average if the price is higher than the average price across all years in this dataset.) In this figure, and throughout the rest of this section, the price for a year is computed according to a price index developed by Ashenfelter. His price index took into account the results of several thousand auction sales for wines from many different wineries in the corresponding year. For more information, see the references in Section 1.5.

Figure 1.1 establishes that it is hot, dry summers that produce vintages in which the mature wines obtain the higher prices. Additionally, the data is fairly consistent; there are very few cases that do not adhere to this rule. This observation led Ashenfelter to build a linear regression model for the price of mature wine as a function of the age of the vintage and the weather. For more about linear regression, see Chapter 21.

A sample of the data he used is in Table 1.1. The regression equation is for the logarithm of the average price of the year, normalized relative to the highest selling year in the data, 1961. For more about why Ashenfelter used the logarithm of price, see Section 1.5. The "Price" variable referenced here is the price index developed by Ashenfelter.

Using this data, Ashenfelter published what is now known as the "Bordeaux equation":

$$\begin{aligned}\text{Log(Price)} = &-12.145 + 0.001173 \times \text{(Winter Rainfall)} \\ &+ 0.616 \times \text{(Average Growing Season Temperature)} \\ &- 0.00386 \times \text{(Harvest Rainfall)} + 0.0238 \times \text{(Age of Vintage)}.\end{aligned}$$

The units for each of the independent variables are given in Table 1.1. This regression equation has an R^2 value of 0.83 and all variables are statistically significant. Ashenfelter experimented with additional variables, but the predictions were remarkably robust to the addition of any other variables.

While this regression equation predicts the logarithm of the average price index of the wine when it is mature, it can be thought of as an equation to predict the quality of the wine. By the time the wine is mature, the quality has been realized, and it is generally believed that the price of the wine reflects the true quality.

In response to Ashenfelter's equation, Britain's *Wine* magazine said "the formula's self-evident silliness invites disrespect." The Bordeaux wine industry was outraged; how can this man claim to know anything about wine that he has never tasted? You would never listen to the opinion of a food critic who had never tasted the food, or to a movie critic who had never seen the movie. To almost everyone, Ashenfelter's approach was just as absurd.

The Weather Makes the Wine

To the surprise of many, Ashenfelter went on to prove that his predictions for the quality of wine are surprisingly accurate. Using his Bordeaux equation, he was able to predict the quality of vintages that had just been bottled and had never even been tasted by comsumers. In 1991, he predicted that the vintages

Figure 1.1: Average growing season temperature and harvest rain in Bordeaux 1952-1980. The years with an above average price index are shown as triangles. The price index was constructed by Ashenfelter to represent the results of several thousand auctions sales for many different wineries in the Bordeaux region.

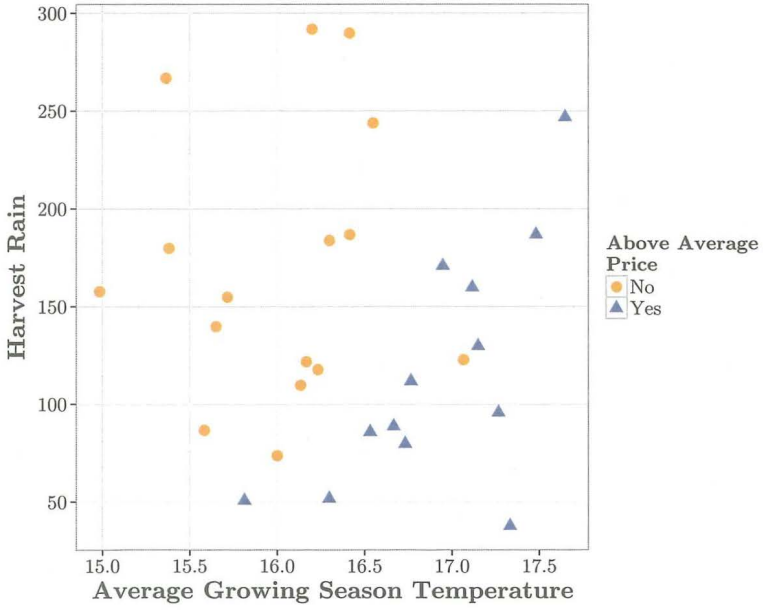


Table 1.1: Vintage and weather data used for the regression equation. The years 1954 and 1956 do not appear here because they were generally considered the worst in their decade and were no longer sold at the time Ashenfelter built his model.

Vintage	Log of Price	Winter Rain (ml)	Average Growing Season Temp (°C)	Harvest Rain (ml)	Age of Vintage (yrs)
1952	-0.99868	600	17.1167	160	31
1953	-0.45440	690	16.7333	80	30
1954		430	15.3833	180	29
1955	-0.80796	502	17.1500	130	28
1956		440	15.6500	140	27
1957	-1.50926	420	16.1333	110	26
1958	-1.71655	582	16.4167	187	25
1959	-0.41800	485	17.4833	187	24
1960	-1.97491	763	16.4167	290	23
1961	0.00000	830	17.3333	38	22
1962	-1.10572	697	16.3000	52	21
1963	-1.78098	608	15.7167	155	20
1964	-1.18435	402	17.2667	96	19
1965	-2.24194	602	15.3667	267	18
1966	-0.74943	819	16.5333	86	17
1967	-1.65388	714	16.2333	118	16
1968	-2.25018	610	16.2000	292	15
1969	-2.14784	575	16.5500	244	14
1970	-0.90544	622	16.6667	89	13
1971	-1.30031	551	16.7667	112	12
1972	-2.28879	536	14.9833	158	11
1973	-1.85700	376	17.0667	123	10
1974	-2.19958	574	16.3000	184	9
1975	-1.20168	572	16.9500	171	8
1976	-1.37264	418	17.6500	247	7
1977	-2.23503	821	15.5833	87	6
1978	-1.30769	763	15.8167	51	5
1979	-1.53960	717	16.1667	122	4
1980	-1.99582	578	16.0000	74	3

of 1989 and 1990 would be exceptional. Many professional wine critics did not agree with him at the time, but there is now a virtually unanimous agreement that 1989 and 1990 are two of the best vintages of the last 50 years. Ashenfelter further predicted that the 2000 and 2003 vintages are in the same league as the 1989 and 1990 vintages. These years have also been praised by the influential Robert Parker who has said "2000 is the greatest vintage Bordeaux has ever

produced." He has equally praised the 2003 vintage. Without tasting a single drop of wine, Ashenfelter was able to come to the same conclusion about the quality of the vintages as the man considered to be the most influential wine critic of our time. Additionally, his method does not require any tasting of the wine; just the use of a simple equation.

1.2 Assessing Quality in Healthcare

Perhaps no other domestic policy topic in the United States has spawned more debate in recent years than healthcare. However, even though there is a significant amount of disagreement about healthcare policies, the ultimate goal of politicians, physicians, hospitals, and patients is the same: *quality* healthcare. But what exactly is quality in healthcare? How is it defined, measured, and improved? One possibility is for an expert physician with many years of experience to look at cases and assess the quality of healthcare that patients have received. Clearly, this is impractical if one wants to make this assessment available for every patient in the healthcare system, as it takes an expert physician approximately an hour to read and understand each case before making an appropriate assessment. Also, how does the physician get the information to make this assessment?

Dimitris Bertsimas and David Czerwinski (two analytics professors), together with Michael Kane (a physician), built a model that assesses the quality of healthcare received by a group of patients. In this work, they define good quality care as healthcare that improves outcomes, educates patients, coordinates care among all doctors that see a patient, and controls costs. The goal is to capture the concerns of the patient, physician, and hospital.

This model is built using the opinion of a physician, and it is intended to accurately predict good or bad quality of care on cases not seen by the physician. With this method, an expert's opinion is used to assess the quality of care, but a model is developed to extend this opinion to all patients, without requiring the expert to evaluate each individual case. For this model, the researchers set the goal of predicting the opinion of Dr. Michael Kane, an internal medicine physician with over 40 years of experience. Keep in mind that the model could easily be extended to predict the average opinion of a committee of physicians, or to predict the opinion of a set of guidelines (this is discussed more later in this section). The key observation here is that the model predicts the opinion of a *domain expert*; this concept could be extended to many other applications and problems.

Claims data, the data that healthcare providers submit to insurance companies to be paid for various services, provides the data for this model. This was the most easily accessible and sufficiently large set of data available electronically and up to date. This data is not 100% accurate, and under-reporting is common, but other data sources are not very accessible. With the increasing use of electronic medical records, there is potential for more accurate

and complete data in future calibrations of the model. We will also see claims data used for different applications later in this book.

The first step in building the model was asking a physician, in this case Dr. Kane, to rate the quality of care of a set of 101 patients. The patients selected were diabetic and between the ages of 35 and 55 with annual healthcare costs between \$10,000 and \$20,000. The creators of the model decided to test it with diabetic patients since there is a wide range of tests, medications, and complications associated with diabetes. The age range was to ensure that the patients should receive similar care (an 18 year old and an 81 year old will probably need very different care, partially due to their age difference), and the cost range was to insure that the patient had enough data to make an accurate assessment, but not so much data that it was impractical for Dr. Kane to review.

Dr. Kane rated the quality of care for each patient on a two point scale: poor care, or good care. He also gave his level of confidence that the patient was indeed receiving that quality of care. These ratings are summarized in Table 1.2. He also wrote a paragraph for each patient, explaining his reasoning. The following paragraph gives an example:

> Male on glucophage, had sporadic medical visits and labs. Did have eye exam 4/05. His primary problems were back pain and narcotic use. He had monthly percocet prescriptions in addition to an NSAID and a muscle relaxer. No diagnostic studies. Had a few physical therapy visits in October and November 2003. No other significant diagnostic or therapeutic initiatives. Poor care with high confidence.

From Dr. Kane's assessments, 80 different variables were defined that fell into six different categories: (1) diabetes treatment (e.g. number of glycated hemoglobin tests); (2) utilization (e.g. number of office visits); (3) markers of good care (e.g. mammogram); (4) markers of poor care (e.g. narcotics); (5) providers (e.g. number of doctors); and (6) prescriptions (e.g. number of different drugs). Since blindly classifying every patient as receiving good care gives an accuracy percentage of 78%, the goal was to be more accurate than this. In addition to accuracy, the types of errors that occur are also significant: the percentage of cases classified as poor care that are actually good care and the percentage of cases classified as good care that are actually poor care. Errors of the first type (good care mistakenly classified as poor care) may cause unnecessary expenses to investigate and treat perfectly fine patients. But errors of the second type (poor care mistakenly classified as good care) can be very dangerous. These errors mean that we are overlooking patients that need help, and might cause serious health issues in the future. Understanding the trade-off between the different types of errors for any model is critical in deciding how useful the model will be in practice.

The model uses logistic regression to predict the binary variable "Quality," which takes value 1 if the patient received good quality care according to the

Table 1.2: The quality of care ratings given by Dr. Kane, along with his level of confidence.

	Low Confidence	High Confidence
Poor Quality	5	17
Good Quality	19	60

assessment of Dr. Kane, and takes value 0 otherwise. Logistic regression is a predictive method used when the outcome variable is binary, and produces a vector of coefficients similarly to linear regression. For more about logistic regression, see Chapter 21.

In this case, the best logistic regression model uses only three different independent variables: a binary variable for whether or not the patient was started on a combination of drugs to treat their diabetes (Started on Combination), a variable for the number of glycated hemoglobin tests the patient was given (Hemoglobin Tests), and a variable for the fraction of time the patient refilled a prescription to treat an acute condition within 30 days of the prescription running out (Acute Refills). The predictive equation is given by:

$$\begin{aligned}\text{Logit}(\text{Quality}) =& 1.66 - 4.23 \times (\text{Started on Combination}) \\ &+ 0.34 \times (\text{Hemoglobin Tests}) - 0.26 \times (\text{Acute Refills}).\end{aligned}$$

This is the *log-odds*, or *logit*, of the model, and is described in more detail in Chapter 21. If we look at this equation, we can see that Started on Combination has a negative coefficient. This means that if everything else is equal, a patient who is started on a combination of drugs (versus a single drug) to treat their diabetes is more likely to have poor quality care. Similarly, Acute Refills has a negative coefficient, so a patient who is repeatedly refilling a drug to treat an acute condition is more likely to have poor quality care. On the other hand, Hemoglobin Tests has a positive coefficient, meaning that more hemoglobin tests is predictive of good quality care. This makes sense, since it is recommended that even healthy diabetics receive this test at least twice a year.

Outcomes

The accuracy of the model for the quality of healthcare is given by the classification matrix in Table 1.3. The rows are labeled by the actual classifications and the columns are labeled by the predicted classifications. Of the cases that were actually classified as poor care, 12 of them were predicted to be poor care and 10 of them were predicted to be good care. Of the cases that were actually classified as good care, 4 of them were predicted to be poor care while 75 of them were predicted to be good care. Approximately half of the poor care patients were classified correctly, and almost all of the good care

patients were classified correctly. This gives an 86% success rate, compared to the baseline of 78%.

To test the model out-of-sample, Dr. Kane rated the care of 30 additional patients that were not used to build the model. The baseline success rate (by blindly rating all patients as good care) in this case was 63%. The accuracy of the model was 80%. This shows that only a simple model is needed to capture half of the cases of poor care.

This model can be used on all patients in the system without having the expert rate all of the patients and classify them accurately. As mentioned earlier, this model could also be extended to predict the quality ratings of a committee of physicians, by having them each rate the quality of care of every patient. The ultimate ratings could then be the most common rating across all physicians. This work shows that a simple model has the potential to be used to replicate the opinion of experts when assessing quality of care.

Table 1.3: The classification matrix for the logistic regression model, comparing the actual outcomes to the predicted outcomes.

	Predicted Poor	**Predicted Good**
Actual Poor	12	10
Actual Good	4	75

1.3 Forecasting Supreme Court Decisions

In many political decisions, predictions are abundant. In the months leading up to a presidential election, predictions are made over and over about who will win. But when it comes to the Supreme Court, predictions are not typically made. However, in 2002, Theodore Ruger, Pauline Kim, Andrew Martin, and Kevin Quinn (two political science professors and two law professors), set out to predict Supreme Court rulings. Most people thought that models did not stand a chance against expert predictions, including the authors themselves. They thought that knowledge of the details of each case and the qualitative aspects of the Court would enable experts to predict much better than any statistical model.

These professors decided to use classification trees as their statistical model, due to the need for a flexible method for pattern detection in a situation with many variables that might not have a linear relationship. For more about classification trees, see Chapter 21.

The ultimate goal was to predict, for each case argued by the Supreme Court in the October 2002 term, whether the Supreme Court Justices would affirm the case (uphold the lower court's decision) or reverse the case (overturn the lower court's decision). The predictions made by the model were only

allowed to depend on six different independent variables: (1) the circuit of origin of the case; (2) the issue area of the case; (3) the type of petitioner; (4) the type of respondent; (5) the ideological direction of the lower court ruling; and (6) whether or not the petitioner argued that a law or practice is unconstitutional. See Table 1.4 for an explanation of the possible responses for each of these variables. These variables were chosen because they are easily observable prior to the oral argument of the case, and they reflect insights into the key factors in previous decisions.

Hoping that the observable patterns in the Justices' past behavior would hold true for their future behavior, the authors used data from 628 cases that had been argued and decided prior to the October 2002 term by this same Supreme Court (the same set of Justices) to train the classification trees. This was a very rare dataset; the Supreme Court Justices are appointed when needed, and so the set of Justices changes at various times. This particular set of Justices had served together for seven years, which was the longest period of time with the same set of Justices in over 180 years.

The final model consists of eleven distinct classification trees. Two of them predict whether the case is likely to be a unanimous "liberal" decision or a unanimous "conservative" decision. These two trees are the first step in the model. If either tree concludes that a unanimous decision is likely, then the process ends. If neither conclude that a unanimous decision is likely (or both, with conflicting results), then nine different trees, one for each Justice, are used to determine the majority ruling. These trees vary significantly from Justice

Table 1.4: Possible values for the variables used in the model. All of the variables used here are categorical variables.

Variable	Possible Responses
Circuit of origin	$1^{st} - 11^{th}$, Federal, D.C.
Issue area of the case	Criminal procedure, civil rights, privacy, etc.
Type of petitioner	An employer, an employee, the United States, etc.
Type of respondent	An employer, an employee, the United States, etc.
Ideological direction of the lower court ruling	Liberal or Conservative
Whether the petitioner argued that a law or practice is unconstitutional	Yes or No

to Justice. For example, Figure 1.2 shows the classification tree for Justice O'Connor.

Figure 1.3 shows the classification tree for Chief Justice Rehnquist (recall that this study was for Justices in the October 2002 term). Note that the classification trees can depend on the decisions of other justices; for example, the tree for Chief Justice Rehnquist depends on the predicted decision for Justice Thomas, requiring that the prediction for Justice Thomas to be made first.

For most people, the simplicity of these classification trees further supported the hypothesis that the experts would do much better than the models. How can it possibly be true that a model using only four yes or no questions can predict the decisions of Justice O'Connor better than experts in the field?

To compare their statistical model to expert opinion, the researchers recruited 83 participants who were considered experts in the field of law. To

Figure 1.2: The classification tree for Justice O'Connor.

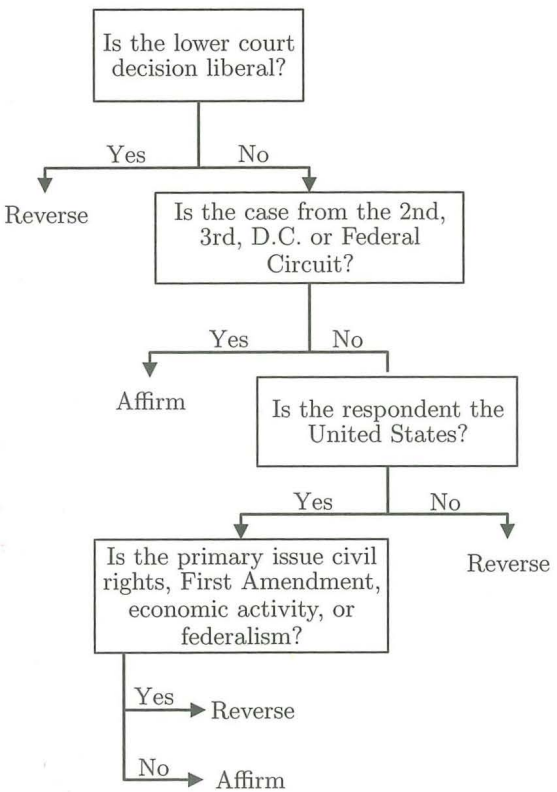
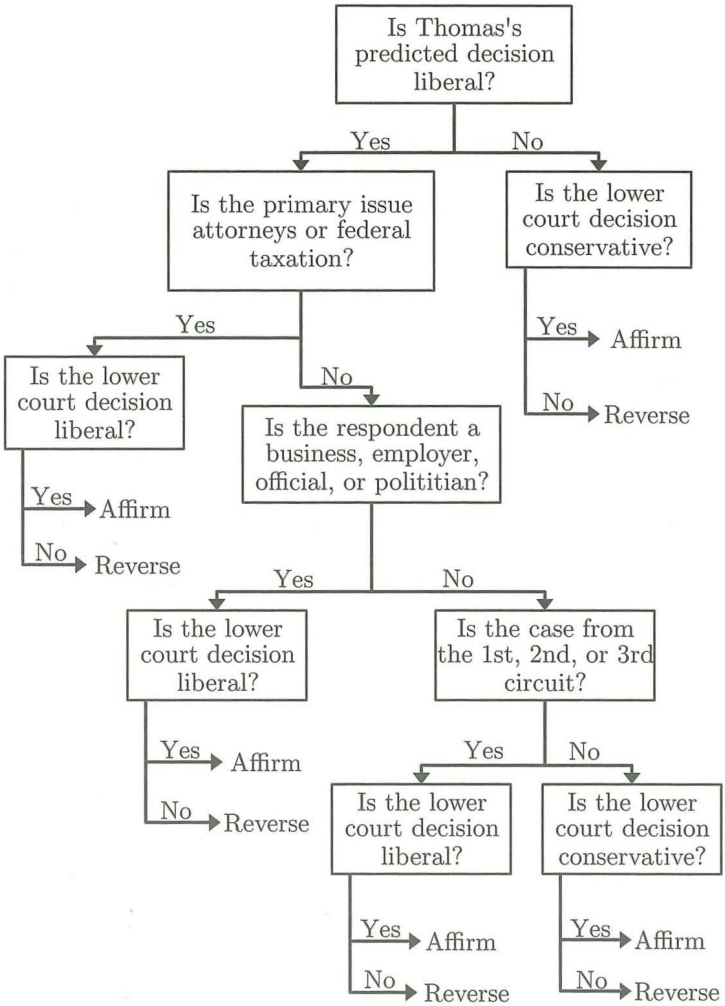


Figure 1.3: The classification tree for Chief Justice Rehnquist.



determine whether someone qualified as an expert, they assessed the person's writings, training and expertise, referrals, and experience with the Supreme Court. The group consisted of 71 academics and twelve appellate attorneys, including 38 former Supreme Court law clerks, 33 chaired professors, and five current or former law school deans. The experts were asked to predict the outcomes of the cases that fell into their area of expertise. More than one expert predicted most cases, and the experts did not communicate about their

predictions. They were asked to make predictions for the Court as a whole and for each individual Justice. They were allowed to use any materials to make their decisions, as well as the specific knowledge that they have gained from years of experience.

Predicting the Unpredictable

When the Supreme Court term started in October 2002, the model had been run and the experts had reached a decision. These predictions were posted publicly on a website, so many people were able to watch with interest as case after case was decided. To the surprise of almost everyone, the model was more accurate at predicting the decisions.

The models and experts were compared using 68 cases for the overall case outcome forecasts (affirm or reject) and 67 of these cases for the individual vote predictions (one case had to be eliminated due to ambiguous individual voting). Since multiple experts assessed the same cases, there were more expert predictions overall (one prediction for each expert and each case). Table 1.5 gives the accuracy of the model and experts in predicting the case outcomes.

The statistical model greatly outperformed the experts in predicting the overall case outcomes (affirm/reverse). An alternative measure of the experts' success takes the majority prediction on each case as the experts' prediction. Even using this more generous measure of the experts' success, their accuracy rate was only 65.6%, still significantly lower than the model's accuracy rate of 75%. These results were shocking; how could such simple models out-predict experts in the field?

The study also compared the success of the model and of the experts in predicting the decisions of each individual justice (Table 1.6). The model and experts were very close in how accurately they predicted the decisions of the individual justices. However, the model and experts were each best at predicting the votes of different Justices. This is shown in Figure 1.4, which compares the proportion of correctly predicted cases by both methods for each individual justice.

The experts were much worse at predicting Justice O'Connor's votes, which is believed to be due to the fact that she is widely viewed as moderate

Table 1.5: The overall accuracy of the model and the experts at predicting the overall case outcomes. Note that since multiple experts assessed each case, there are more expert predictions overall.

	Correct	Incorrect
Model	51 (75.0%)	17 (25.0%)
Experts	101 (59.1%)	70 (40.9%)

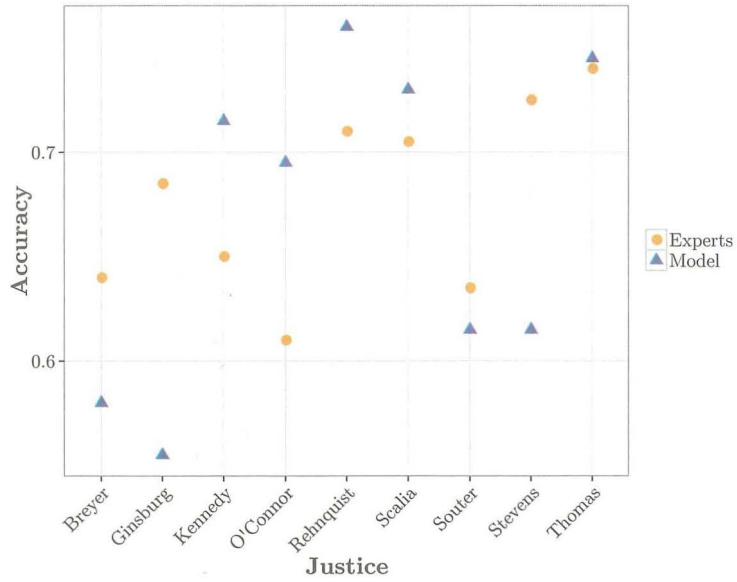
by observers of the Court. However, the statistical model was able to correctly predict her votes much more frequently. This shows that the model was able to capture patterns in her decision making that were not obvious to the experts. On the other hand, the experts were able to more accurately predict the votes of other justices, in particular those that are thought of as more conservative.

The success of the model to predict the more "unpredictable" justice votes confirms the belief that models are able to see patterns in data that humans have a harder time discovering. However, the success of the experts in predicting the more "predictable" justice votes shows that intuition and human reasoning of experts can be a better tool in certain cases. This example shows that a mixture of models and experts might be the best approach; the unemotional

Table 1.6: The overall accuracy of the model and the experts at predicting the decisions of the individual justices. Note that some justices did not vote on some cases, and are therefore excluded from this analysis.

	Correct	Incorrect
Model	400 (66.7%)	200 (33.3%)
Experts	1015 (67.9%)	479 (32.1%)

Figure 1.4: Model and expert accuracy percentages for individual Justices.



and consistent nature of a model mixed with the intuition and knowledge of an expert gives a much stronger prediction than either of the two alone.

1.4 The Competitive Edge of Models

These three examples provide insight on using models to make predictive decisions. Most importantly, we have learned that:

1. A model does not replace an expert. Expert human judgment is needed to develop a model, interpret the results, and adjust the model when necessary.

2. To make predictions on a large scale, the only option is a model. This was very clear in the healthcare example, since it is impossible for a physician to read and understand millions of patient records before making an accurate assessment.

3. In many cases, models can provide an edge over expert human judgment. People are very good at understanding and analyzing small amounts of information, but good models are able to look at all of the known information quickly and accurately. Models can also aggregate the opinions and assessments of many people into one final unbiased and unemotional prediction. A good model allows the expert to make a more informed decision, as well as gives the expert more time to work on other tasks that a model or computer can not perform as well.

The use of models to make predictions that have typically been made by experts has been a source of controversy; the experts do not like the idea of feeling unnecessary, and in general, people are uneasy with listening to the opinion of a computer program rather than the opinion of an expert in the field. But these examples have shown that the model is not replacing the expert; the model is just performing tasks that can be more efficiently and accurately performed by a computer.

1.5 Notes and Sources

1.1. The information on the Bordeaux equation comes from the academic article [7] and from the *New York Times* article "Wine equation puts some noses out of joint" [113]. The data presented here can be found on the *Liquid Assets* webpage, the wine journal published by Ashenfelter [8].

In this model, Ashenfelter chose to predict the logarithm of price, instead of directly predicting the price of the wine. This is a common practice in the field of economics, and captures what is called the *real*

product rate of return. A useful property of the natural logarithm is that it converts an exponential growth pattern to a linear growth pattern. Due to inflation and potentially other factors, the price of wine is expected to increase exponentially over time. By taking the logarithm, the dependent variable (price) shows a more linear pattern from year to year, which makes it more suitable for predicting with a linear regression model.

1.2. The study assessing the quality of care in the United States is published in *The Milbank Quarterly* [125]. The recommended actions for diabetic care are from the American Diabetes Association [3]. The model for assessing the quality of care is described in more detail in the doctorate thesis of David Czerwinski [39] and in the paper by Bertsimas, Czerwinski, and Kane [18].

1.3. The Supreme Court study is described in the academic papers [117, 118]. The figures and diagrams here are reproduced from the results in these papers.