DS Baseline

Ivan Bizberg

2023-07-17

Table of contents

Preface

Welcome to the world of data science, a captivating realm where art, mathematics, and technology converge to unlock the hidden insights lying dormant within vast troves of information. In this book, we embark on an exhilarating journey through the most important aspects of data science techniques, exploring their foundations, applications, and transformative potential.

The rapid advancement of technology and the explosive growth of data have ushered in an era of unprecedented opportunities. Every click, every transaction, and every interaction generates a data footprint that can be harnessed to drive innovation, solve complex problems, and shape the future. Data science equips us with the tools to make sense of this digital universe, enabling us to extract meaning from seemingly chaotic data and turn it into actionable knowledge.

At its core, data science is a multidisciplinary field that draws from various domains, including statistics, mathematics, computer science, and domain expertise. It encompasses a wide range of techniques, algorithms, and methodologies designed to collect, analyze, interpret, and visualize data to uncover patterns, make predictions, and generate insights.

This book is intended to serve as a comprehensive compilation of the most important aspects of data science techniques. However, it is important to note that the content presented herein reflects the author's personal understanding and interpretation of these concepts which may occasionally deviate from the precise and mathematical definitions of certain data science principles.

1 Terms & definitions

1.1 Correlation vs covariance

- **Covariance** measures the direction and magnitude of the linear relationship between two variables. It calculates how changes in one variable are related to changes in another variable. Covariance can take any value, positive or negative, depending on the nature of the relationship.

When to use it:

- Scaling and Interpretation: If you are primarily interested in the magnitude of the relationship between two variables, without the need for standardized interpretation, covariance can be used. Since covariance is not standardized, it preserves the original scale of the variables. This can be helpful when the units of measurement carry important information or when you want to maintain the original context of the data.
- Non-linear Relationships: If you suspect a non-linear relationship between variables, covariance can still provide insights into the direction and magnitude of the relationship, albeit without quantifying the strength in a standardized manner.

Example:

By calculating the covariance between height and weight, you can obtain a measure of how the two variables vary together.

- **Correlation** measures the strength and direction of the linear relationship between two variables, but it standardizes the measure to fall between -1 and 1.

When to use it: Correlation is a more useful measure than covariance when comparing relationships across different datasets or variables, as it removes the influence of the scales of the variables.

Example:

If you were interested in comparing the relationship between height and weight with other datasets or variables, or if you wanted a standardized measure of the strength of the relationship, then calculating the correlation coefficient would be more appropriate. The correlation coefficient would provide a standardized measure between -1 and 1, allowing for easier comparison and interpretation across different contexts.

1.2 Correlation vs cointegration

Both are statistical concepts used to measure the relationship between variables.

- **Cointegration** measures whether the variables tend to move together over time, despite possibly having short-term fluctuations.

Example: A drunk man leaves the pub with his dog.

When the man and the dog first leave the pub, their paths are **correlated**. They generally move in the same direction, but the distance between the dog and the man has no actual limit. It increases at times, decreases at times, but is generally random and poorly defined. The direction of the two, however, is generally the same.

When the man leashes his dog to cross the road, they become **cointegrated**. Now, while their direction is still the same, their distance from one another is finite. The dog cannot move beyond the length of the leash from the man. ("What Is the Difference Between Correlation and Cointegration? Is Cointegration a Good Measure of Risk?" n.d.)

1.3 P-value

A p-value of 0.001 indicates that if the null hypothesis tested were indeed true, then there would be a one-in-1,000 chance of observing results at least as extreme.

1.4 Bias and Variance

We search a model with low bias and low variance

Bias: The inability for a machine learning method to capture the true relationship is called bias **Variance**: The difference in fits between train set and test set

Red model: low variance: This model has low variance if a new point it would not substantially change the model fit (leads to under-fitting) high bias: Ineffective at modeling the data

Blue model: high variance: Small perturbations in the data will significantly change the model fit (leads to over-fitting) low bias: More complex and flexible allowing to model very good the data

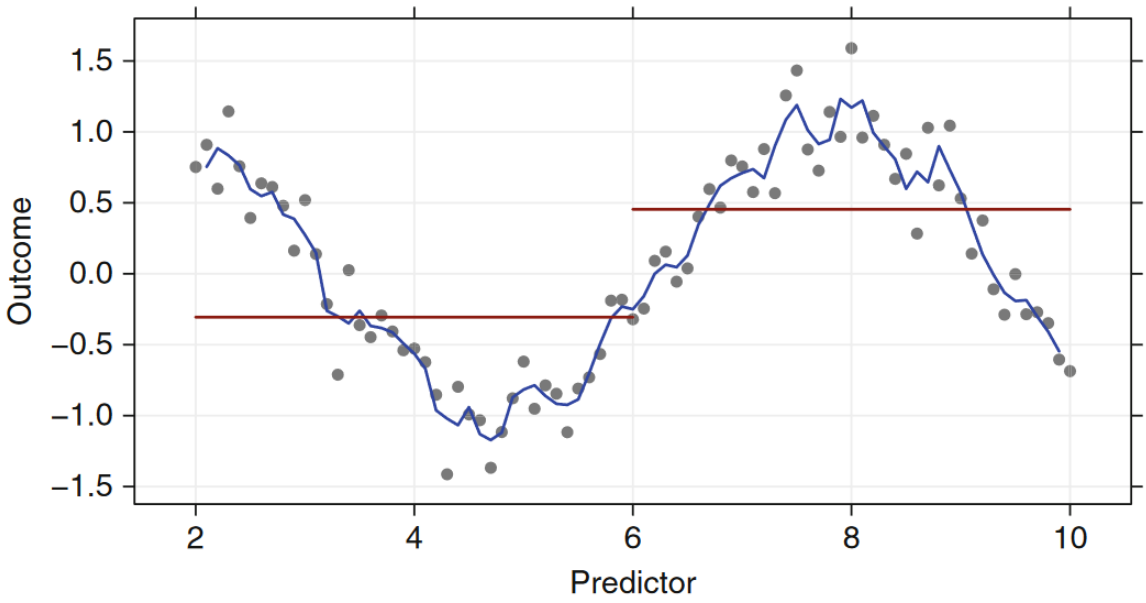

Figure 1.1: variance-bias trade-off

1.5 Entropy

1.6 Regularization

It make the prediction less sensitive to the training data.

1.7 Quantiles

Median: The 50th percentile, which divides the data into two equal halves. Half of the data points are below the median, and half are above it.

Quartiles: The 25th, 50th, and 75th percentiles are called the first quartile (Q1), the median (Q2), and the third quartile (Q3), respectively. They divide the data into four equal parts, each containing 25% of the data.

Percentiles: These are any values that divide the data into 100 equal parts. For example, the 20th percentile is the value below which 20% of the data falls, and the 80th percentile is the value below which 80% of the data falls.

2 Greedy algorithms

When a procedure is greedy, it means that it does not reevaluate past solutions.

3 Extrapolation

Extrapolation is commonly defined as using a model to predict samples that are outside the range of the training data. To know if we can trust our model we need to compare the predictor space between the training data and new data.

- This can be done using **PCA**. If the training data and new data are generated from the same mechanism, then the projection of these data will overlap in the scatter plot. However, if the training data and new data occupy different parts of the scatter plot, then the data may not be generated by the same mechanism and predictions for the new data should be used with caution.

- This can also be done using the following **algorithm**

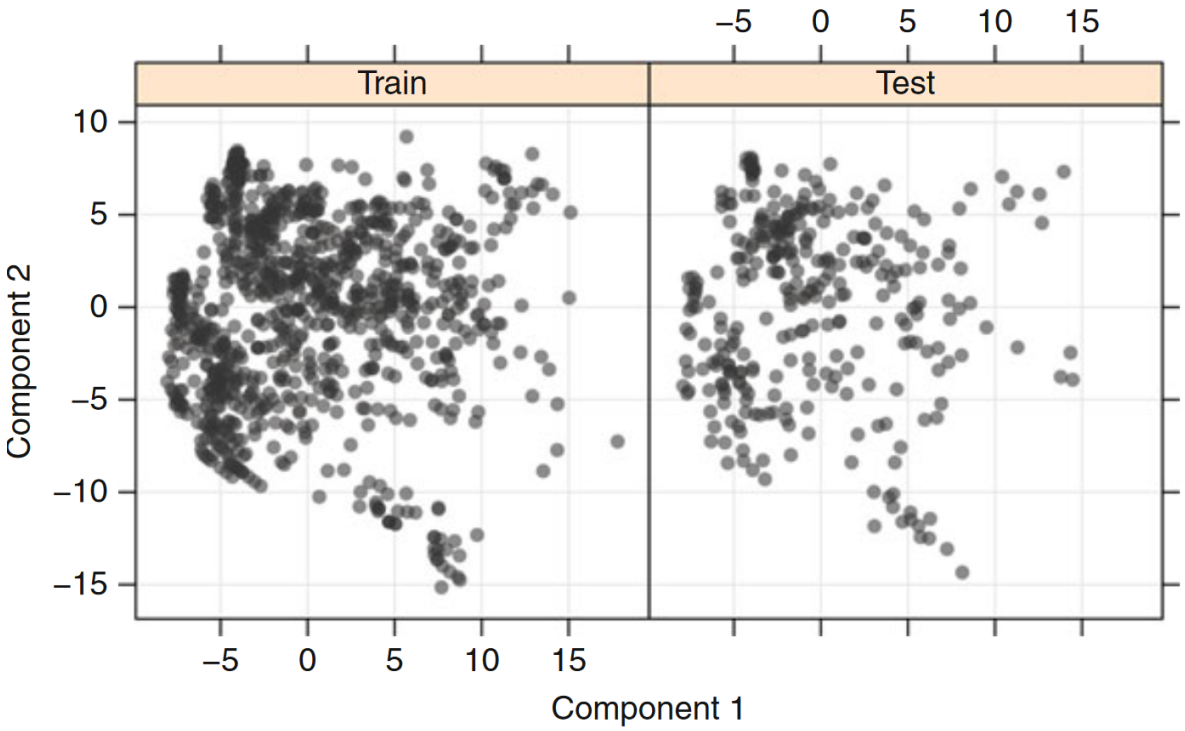

Fig. 20.7: PCA plots for the solubility data

Figure 3.1: PCA and Extrapolation: the training and testing data appear to occupy the same space as determined by these components.

1. Compute the variable importance for the original model and identify the top 20 predictors
2. Randomly permute these predictors from the training set
3. Row-wise concatenate the original training set's top predictors and the randomly permuted version of these predictors
4. Create a classification vector that identifies the rows of the original training set and the rows of the permuted training set
5. Train a classification model on the newly created data
6. Use the classification model to predict the probability of new data being in the class of the training set.

Algorithm 20.1: Algorithm for determining similarity to the training set

Figure 3.2: Algorithm and Extrapolation

4 Choose machine learning model

Model	Mode	Allows n < p	
Linear regression	regression	no	centering and scaling, remove near-zer
Partial least squares	regression	depends	cente
Logistic regression	classification	no	centering and scaling, remove near-zer
Ridge regression	regression & classification	no	centering and scalin
Elastic net/lasso	regression & classification	yes	ce
Support vector machines	regression & classification	yes	cente
MARS/FDA	regression & classification	yes	
K-nearest neighbors	regression & classification	yes	ce
Random forest	regression & classification	yes	
Boosted trees	regression & classification	yes	
Neural networks	regression & classification	yes	centering and scaling, remove near-zer
LDA	classification	no	
Naive Bayes	classification	yes	
C5.0	classification	yes	

(Kuhn and Johnson 2013)

Common steps in model building - Pre-processing the predictor data - Estimating model parameters - Selecting predictors for the model - Evaluating model performance - Fine tuning class prediction rules (via ROC curves, etc.)

- optimization routines (e.g., Nelder–Mead simplex method = *direct methods*) can later be use to search the optimal key value (e.g., determine possible mixtures with improved compressive strength)

5 Pre-processing

5.1 Transform the predictor/independent variables

Important to transform independent variables that are skewn or containt outliers for models sensitive to them

5.1.1 Centering and scaling

Advantages: Improve the numerical stability to minimize potential numerical errors **Disadvantage**: Loss of interpretability

5.1.2 Resolve Skewness

Skewed data: Ratio of the highest value to the lowest value is greater than 20 have significant skewness.

- log
- square root
- inverse
- Box and Cox

5.1.3 Resolve Outliers

- Spatial sign

NOTES:

- it is important to center and scale the predictor data prior to using this transformation
- spatial sign transformation of the predictors transforms them as a group. Removing predictor variables after applying this technique may be problematic.

5.1.4 Data Reduction and Signal/Feature Extraction

These methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables.

5.1.4.1 PCA

The number of components to retain is choosen by creating a scree plot (Fig 1)

For most data sets, the first few PCs will summarize a majority of the variability, and the plot will show a steep descent; variation will then taper off for the remaining components. Generally, the component number prior to the tapering off of variation is the maximal component that is retained. In an automated model building process, the optimal number of components can be determined by cross-validation (see Resampling Techniques).

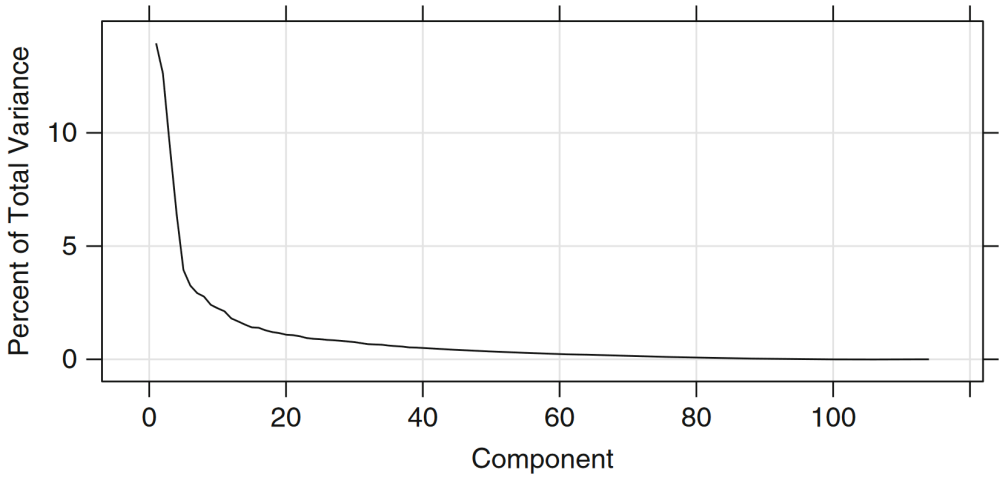

Fig. 3.6: A "scree plot" where the percentage of the total variance explained by each component is shown

Figure 5.1: **Figure 1**: The variation tapers off at component 5. Using this rule of thumb, four PCs would be retained

Advantages: The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated.

Disadvantage:

- Loss of interpretability. PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective.

- Unsupervised technique which means that PCA it does not consider the modeling objective or response variable when summarizing variability. If the predictive relationship between the predictors and response is not connected to the predictors' variability, then the derived PCs will not provide a suitable relationship with the response. In this case, a supervised technique, like PLS, will derive components while simultaneously considering the corresponding response.

NOTES:

- first transform skewed predictors and then center and scale the predictors prior to performing PCA. Centering and scaling enables PCA to find the underlying relationships in the data without being influenced by the original measurement scales.
- If PCA has captured a sufficient amount of information in the data. Visually examining the principal components is a critical step for assessing data quality and gaining intuition for the problem. To do this, the first few principal components can be plotted against each other and the plot symbols can be colored by relevant characteristics, such as the class labels.
  - Check for blatant outliers that may prompt a closer examination of the individual data points
  - Check for clusters of samples (for classification problems; Try other models that could better accommodate the data to have a final conclusion)
  - Checks loadings to characterize which predictors are associated with each component (Loadings close to zero indicate that the predictor variable did not contribute much to that component; Fig 2)
  - Check for multicollinearity (substantial correlation between multiple predictors): For example, if the first principal component accounts for a large percentage of the variance, this implies that there is at least one group of predictors that represent the same information. For example, Fig 1 indicates that the first 3–4 components have relative contributions to the total variance. This would indicate that there are at least 3–4 significant relationships between the predictors. Colinearity can increase the model variance
- If the percentages of variation explained are not large (e.g., less than 48 %) for the first three components, it is important not to over-interpret the resulting image.

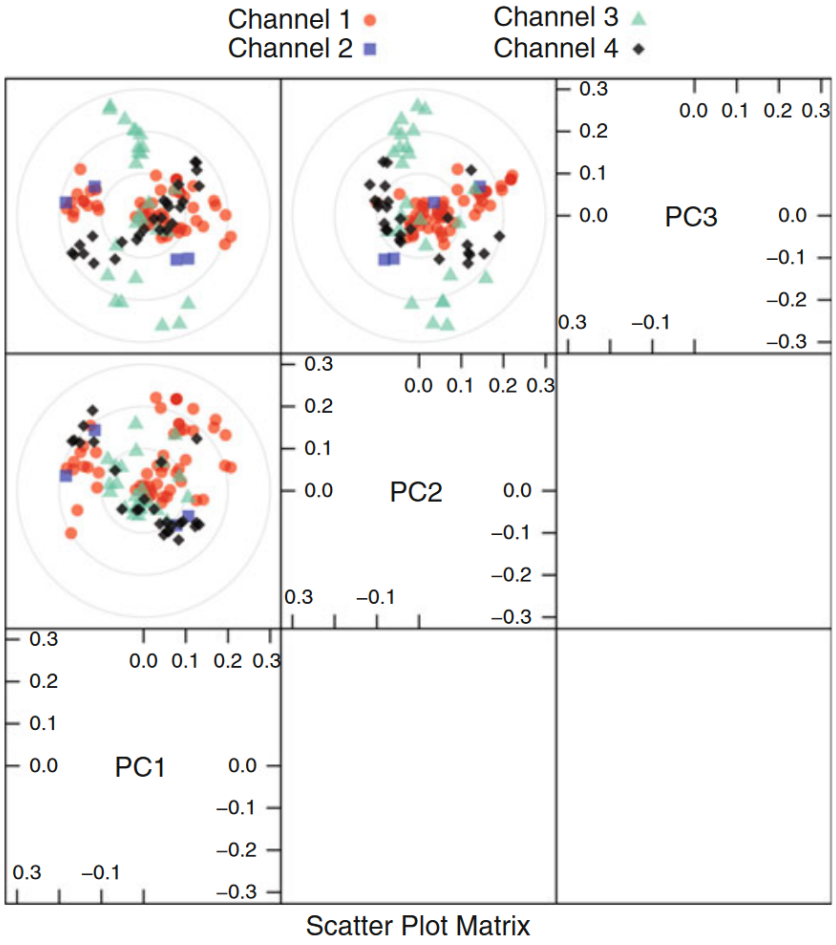

Figure 5.2: **Figure 2**: loadings for the first three components in the cell segmentation data. For the first principal component, the loadings for the first channel are on the extremes. This indicates that channel 1 have the largest effect on the first principal component and by extension the predictor values. Also note that the majority of the loadings for the third channel are closer to zero for the first component. Conversely, the third principal component is mostly associated with the third channel while the first channel plays a minor role here.

5.1.4.2 PLS (*to do*)

5.1.5 Dealing with Missing Values

Before proceeding: It is important to understand why the values are missing to check for *informative missing*. Informative missingness can induce significant bias in the model.

Examples: - people are more compelled to rate products when they have strong opinions (good or bad) - The tested drug was extremely ineffective or had significant side effects. The patient may be likely to miss doctor visits or to drop out of the study.

- **1) Remove predictors** from models if the percentage of missing data is substantial
- **2)** For large data sets, **removal of samples** based on missing values is not a problem
- **3** If we **do not remove** the missing data:
  - Use tree-based techniques, which account for missing data
  - impute missing data using information in the training set predictors to estimate the values of other predictors. This amounts to a predictive model within a predictive model. If we are using resampling to select tuning parameter values or to estimate performance, the imputation should be incorporated within the resampling. This will increase the computational time for building models, but it will also provide honest estimates of model performance.
    - K-nearest neighbor model:

    Advantages: The imputed data are confined to be within the range of the training set values. Disadvantages: The entire training set is required every time a missing value needs to be imputed **and** The number of neighbors is a tuning parameter
    - Linear regression model: between a predictor with few missing points strongly associated with the predictor with missing data (correlation / visualizations / PCA.

NOTES Censored data Missing data: The exact value is missing but something is known about its value. For example: If a customer has not yet returned a movie to blockbuster, we do not know the actual time span, only that it is as least as long as the current duration.

- For inference models: the censoring is usually taken into account in a formal manner by making assumptions about the censoring mechanism
- For predictive models, it is more common to treat these data as simple missing data or use the censored value as the observed value.

5.1.6 Removing Predictors

Advantages: Does not compromise the performance and stability of the model. Decreased computational time and complexity. Lead to a more parsimonious and interpretable model

- Remove near-zero predictors (e.g., predictor variable where the percentage of unique values is low $< 10\%$ = unique values/total values and The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large > 20).
- Remove problematic predictors with degenerate distributions as some models can be crippled by them
- Remove highly correlated predictors as both mesure the same underlaying information. **For linear regressions use VIF for other models ensure that all pairwise correlations are below 0.75 threshold for sensitive models**

5.1.7 Adding Predictors (Feature engineering)

- Dummy variables: If the model allow it, dummy variables would help improve interpretation of the model.
- Squared predictors for simplistic models such as regressions
- class centroids for classification models: centers of the predictor data for each class. For each predictor, the distance to each class centroid can be calculated and these distances can be added to the model.
- When the response is bimodal (or multimodal), categorizing the response is appropriate.

AVOID: Binning Predictors (e.g., Temperature less than 36 $\circ$C or greater than 38 $\circ$C.)

6 Model Tuning

Identify settings for the model's parameters that yield the best and most realistic predictive performance

6.1 Data splitting method

A good rule of thumb is about 75–80 % on train subset and the rest for the test subset. Proportionally large test sets divide the data in a way that increases bias in the performance estimates.

6.1.1 Nonrandom approaches to splitting the data

Example: - If a model was being used to predict patient outcomes, the model may be created using certain patient sets (e.g., from the same clinical site or disease stage), and then tested on a different sample population to understand how well the model generalizes.

- In chemical modeling for drug discovery, new "chemical space" is constantly being explored. We are most interested in accurate predictions in the chemical space that is currently being investigated rather than the space that was evaluated years prior.
- In spam filtering; it is more important for the model to catch the new spamming techniques rather than prior spamming schemes.

6.1.2 Random sampling methods

6.1.2.1 Simple random sample

- The simplest way to split the data randomly into a training and test.

  Disadvantage: limited ability to characterize the uncertainty in the results.

- simple k-Fold Cross-Validation: The samples are randomly partitioned into k sets of roughly equal size. A model is fit using the all samples except one subset. The held-out samples are predicted by this model and used to estimate performance measures. The first subset is returned to the training set an procedure repeats with the next subset held out, and so on. Performance estimates, are calculated from each set of held-out samples and then averaged.

  NOTES: The choice of k is usually **5** or **10**, but there is no formal rule. **The bias is smaller for k = 10 than k = 5.** As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. But larger values of k are more computationally burdensome.

  Advantage: Low computational costs. . **Disadvantage**: k-fold cross-validation generally has high variance compared to other methods (only for small training sets). **USE**: If sample sizes are large ($> 10\ 000$) and we want to choose tuning parameters

- repeated k-Fold Cross-Validation

  Advantage: Increase the precision of the estimates while still maintaining a small bias. The bias and variance properties are good and, given the sample size, the computational costs are not large. **Disdvantage**: Large computational costs. **USE**: with k = 10; If the samples size is small (< 1000 obs) and we want to choose tuning parameters

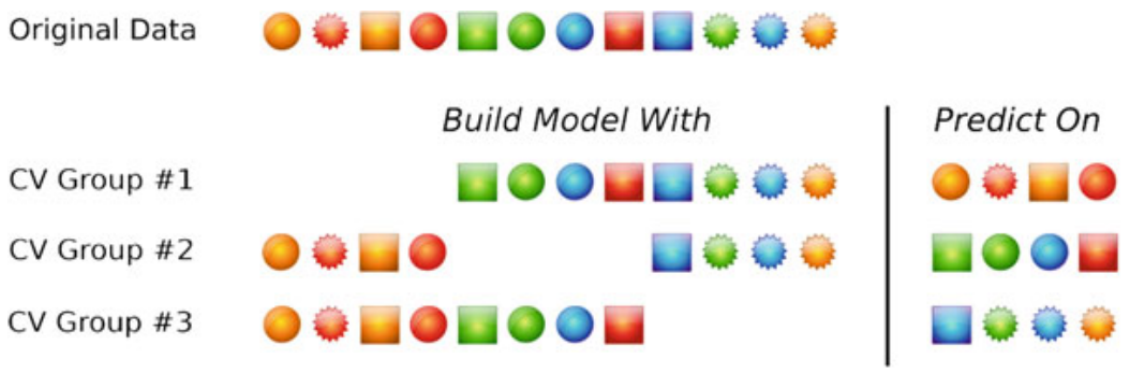

Figure 6.1: k-Fold Cross-Validation

- leave-one-out Cross-Validation / LOOCV: fits as many models as there are samples in the training set, should only be considered when the number of samples is very small.
- leave-group-out Cross-Validation / Repeated training/test splits / Monte Carlo cross-validation: Same as k-fold cross-validation except that samples can be represented in multiple held-out subsets. Also, the number of repetitions is usually larger than in k-fold cross-validation

  Disadvantage:

NOTES: Increase the number of repetition can allow to increase the proportion of data in the train set and decreasing the uncertainty of the performance estimates. To get stable estimates of performance, it is suggested to choose a larger number of repetitions **(say 50–200)**

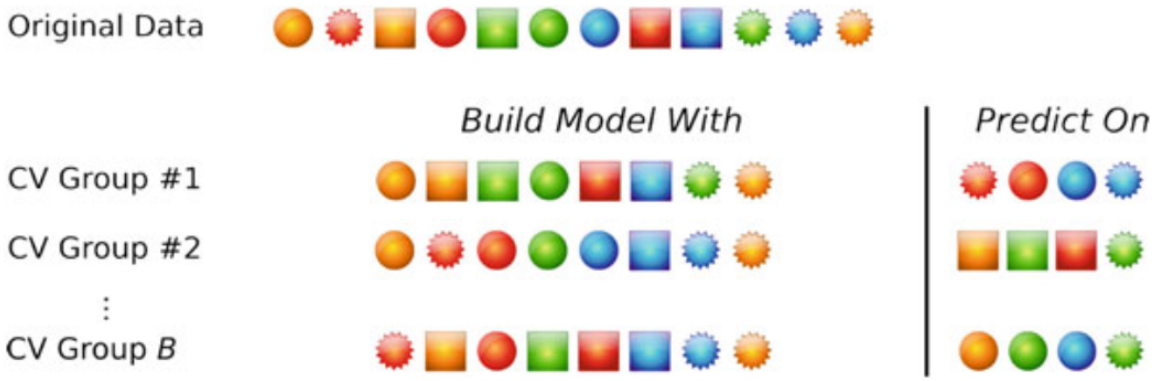

Figure 6.2: leave-group-out Cross-Validation

- The Bootstrap: Each train subset is the same size as the original and can contain multiple instances of the same data point (taken with replacement). Samples not selected by the bootstrap ("out-of-bag" samples) are predicted and used to estimate model performance

  Advantage: error rates have less uncertainty than k-fold cross-validation. Very low variance. **Disadvantage**: On average, 63.2 % of the data points the bootstrap sample are represented at least once, so this technique has bias. similar to k-fold cross-validation when k 2. If the training set size is small, this bias may be problematic, but will decrease as the training set sample size becomes larger. **USE**: If the goal is to choose between models (boosted trees vs support vector machines…), as opposed to getting the best indicator of performance

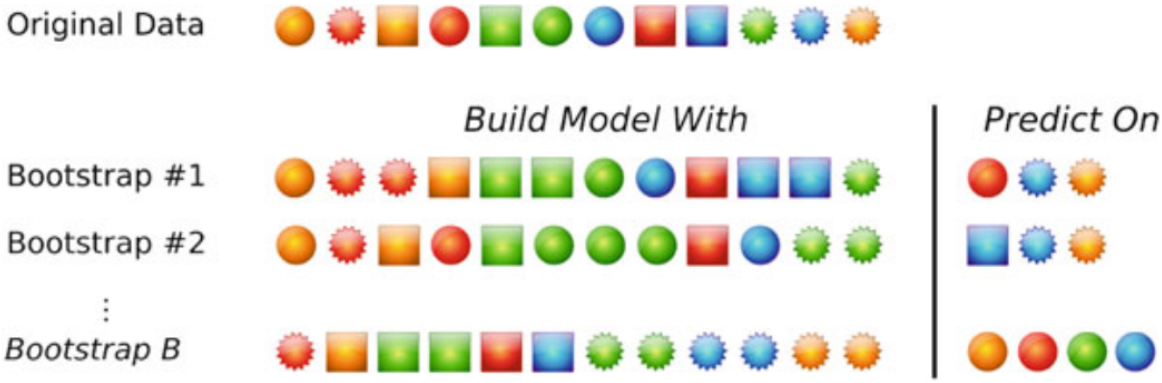

Figure 6.3: Bootstrap

- The Bootstrap 632 method

  Advantage: The modified bootstrap estimate reduces the bias.

  Disadvantage: The estimate is unstable with small samples sizes. This estimate can also result in unduly optimistic results when the model severely over-fits the data, since the apparent error rate will be close to zero.

- The Bootstrap 632+ method **Advantage**: Allows to adjust the bootstrap 632 method estimates

6.1.2.2 Stratified random

To account for the outcome when splitting the data. Applies random sampling within sub-groups (such as the classes or is outcomes are numbers the numeric values are broken into similar groups (e.g., low, medium, and high)).

- k-Fold Cross-Validation

6.1.2.3 Maximum dissimilarity sampling

The data is split on the basis of the predictor values.

6.2 Choosing Tuning Parameters

- Pick the settings associated with the numerically best performance estimates. **Disadvantage**: lead to models that are overly complicated
- Pick simpler models that provide acceptable performance (relative to the numerically optimal settings)
  - The "one-standard error" method: pick the simplier model within a single standard error of the numerically best value. In table below we would pick cost value of 2.
  - the "percent decrease in performance" method: pick the simplier model that is within a certain tolerance of the numerically best value. (e.g., The percent decrease in performance could be quantified by $(X - O)/O$ where X is the performance value and O is the numerically optimal value. For example, in Fig. 4.9, the best accuracy value across the profile was 75 %. If a 4 % loss in accuracy was acceptable as a trade-off for a simpler model, accuracy values greater than 71.2 % would be acceptable. For the profile in Fig. 4.9, a cost value of 1 would be chosen using this approach.)

Resampled accuracy (%)			
Cost	Mean	Std. error	% Tolerance
0.25	70.0	0.0	−6.67
0.50	71.3	0.2	−4.90
1.00	74.0	0.5	−1.33
2.00	74.5	0.7	−0.63
4.00	74.1	0.7	−1.20
8.00	75.0	0.7	0.00
16.00	74.9	0.8	−0.13
32.00	72.5	0.7	−3.40
64.00	72.0	0.8	−4.07
128.00	72.0	0.8	−4.07

Figure 6.4: Cross-validation accuracy

6.3 Metrics / performance measures

6.3.1 For models predicting a categorical outcome

6.3.1.1 Accuracy based metrics

A good model has generally a metric above 0.7 / 70%

1) **Accuracy**:

- *Higher Value*: Better performance (values form 0 to 1)
- *When to use*: Use when the classes are balanced (e.g., Suppose the rate of this disorder1 in fetuses is approximately 1 in 800 or about one-tenth of one percent. A predictive model can achieve almost perfect accuracy by predicting all samples to be negative for Down syndrome.), and misclassification of different classes has similar consequences.
- *Advantage*: Simple and easy to interpret.
- *Disadvantage*: Can be misleading when classes are imbalanced / make no distinction about the type of errors being made
- *Description*: Accuracy measures the proportion of correct predictions out of all predictions made by the model.

- *Example*: Suppose you have a binary classification problem to identify whether an email is spam or not. If your model has an accuracy of 90%, it means it correctly classified 90% of the emails.
- *Calculation*: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives) (Number of Correct Predictions) / (Total Number of Predictions)
- *Notes*:
  - When evaluating the accuracy of a model, the baseline accuracy rate to beat would be the percentage which could be achieve by simply predicting all samples to the dominant category (e.g., In the data set, 70 % were rated as having good, accuracy rate to beat would be 70 % which is the no-information rate).
  - **error rate**: (Number of Incorrect Predictions) / (Total Number of Predictions)

1.5) **Kappa**:

- *Higher Value*: Better performance (values from -1 to 1; 0.30 to 0.50 indicate reasonable agreement)
- *When to use*: Rather than calculate the overall accuracy and compare it to the no-information rate, *Kappa* can be used that take into account the class distributions of the training set samples.
- *Advantage*: Takes into account the accuracy that would be generated simply by chance.
- *Disadvantage*: NA
- *Description*: assess the agreement between two raters
- *Example*: 0 means there is no agreement between the observed and predicted classes, while a value of 1 indicates perfect concordance of the model prediction and the observed classes. Negative values indicate that the prediction is in the opposite direction of the truth, but large negative values seldom occur, if ever, when working with predictive models.
- *Calculation*: Kappa = O − E / 1−E: O is the observed accuracy and E is the expected accuracy based on the marginal totals of the confusion matrix.
- *Note*: The Kappa statistic can also be extended to evaluate concordance in problems with more than two classes. When there is a natural ordering to the classes (e.g., "low,""medium," and "high"), an alternate form of the statistic called weighted Kappa can be used to enact more substantial penalties on errors that are further away from the true result. For example, a "low" sample erroneously predicted as "high" would reduce the Kappa statistic more than an error were "low" was predicted to be "medium." See (Agresti 2002)for more details.

2) **Precision**:

- *Higher Value*: Better performance (good = 0.7)

- *When to use*: Use when the cost of false positives is high (e.g., medical diagnosis, fraud detection).
- *Advantage*: Focuses on the relevance of positive predictions.
- *Disadvantage*: Ignores true negatives and may not be suitable for imbalanced datasets.
- *Description*: Precision is the proportion of true positive predictions (correctly predicted positive class) out of all positive predictions made by the model.
- *Example*: In the spam email example, if your model has a precision of 80%, it means that out of all the emails it predicted as spam, 80% of them were actually spam.
- *Calculation*: True Positives / (True Positives + False Positives)

3) **Sensitivity / True positive rate / Recall**:

- *Higher Value*: Better performance
- *When to use*: Use when the cost of false negatives is high (e.g., medical diagnosis, safety-critical applications).
- *Advantage*: Focuses on the completeness of positive predictions (includes true positive and false negatives).
- *Disadvantage*: Ignores true negatives. For many classification problems, sensitivity may be misleading specially under class imbalance. Since a better cutoff may be possible, an analysis of the ROC curve can lead to improvements in these metrics. Consequently, performance metrics that are independent of probability cutoffs are likely to produce more meaningful contrasts between models.
- *Description*: Is the proportion of true positive predictions out of all actual positive instances in the dataset.
- *Example*: % of people *with* heart diseases were correctly identify by the model
- *Calculation*: TP / (TP + FN)
- *Notes*:
  - If the data set includes more events than nonevents, the sensitivity can be estimated with greater precision than the specificity and sensitivity shouls be use to choose between models.
  - When we want to make unconditional evaluations of the data: know for example what are the chances that … (e.g., If PPV = 0.75 this means that out of all the individuals who tested positive for Disease X, 75% of them actually have the disease, while the remaining 25% are false positives) we can use positive predicted value (PPV = Sensitivity × Prevalence / (Sensitivity × Prevalence) + ((1 − Specificity) × (1 − Prevalence))) IMPOTANT: Predictive values are not often used to characterize the model. There are several reasons why, most of which are related to prevalence. First, prevalence is hard to quantify.

4) **Specificity / True Negative Rate**:

- *Higher Value*: Better performance

- *When to use*: Use when you want to focus on correctly identifying negative cases and the cost of false positives is high.
- *Advantage*: Focuses on the negative class and avoids false positives. Can be misleading specially under class imbalance.
- *Disadvantage*: Ignores true positives.
- *Description*: Specificity measures the proportion of true negative predictions out of all actual negative samples.
- *Example*: % of people *without* heart diseases were correctly identify by the model
- *Calculation*: True Negatives / (True Negatives + False Positives);
- *Notes*:
  - When we want to make unconditional evaluations of the data: know for example what are the chances that … (If NPV = 0.966, this means that out of all the individuals who tested negative for Disease X, 96.6% of them truly do not have the disease, while the remaining 3.4% are false negatives (individuals who have the disease but were incorrectly identified as negative).) we can use negative predicted value (NPV = Specificity × (1 − Prevalence) / (Prevalence × (1 − Sensitivity)) + (Specificity × (1 − Prevalence))). IMPORTANT: *idem*
  - *False-positive rate* : one minus the specificity

4.3) **Youden's J Index**

- *Higher Value*: Better performance
- *When to use*: Use when you want a measure that reflects the false-positive and false-negative rates and summarize the magnitude of both types of errors.
- *Advantage*: Focuses on the negative class and avoids false positives.
- *Disadvantage*: Ignores true positives and may not be suitable for imbalanced datasets.
- *Description*: measures the proportions of correctly predicted samples for both the event and nonevent groups.
- *Example*: % of people *without* heart diseases were correctly identify by the model
- *Calculation*: J = Sensitivity + Specificity − 1

5) **F1 Score**:

- *Higher Value*: Better performance
- *When to use*: Use when there is a trade-off between precision and recall.
- *Advantage*: Incorporates both precision and recall into a single metric.
- *Disadvantage*: Ignores true negatives, which can be important in some cases. May not be ideal for highly imbalanced datasets.
- *Description*: F1 score is the harmonic mean of precision and recall, providing a balance between the two.
- *Example*: Let's say your model has an F1 Score of 0.75, it means there is a balanced trade-off between correctly identifying positive samples and minimizing false positives.
- *Calculation*: 2 * (Precision * Recall) / (Precision + Recall)

6.3.1.2 Class probabilities

Class probabilities potentially offer more information about model predictions than the simple class value. This

4.7) **ROC**:

- *Higher Value*: Better performance (A perfect model that completely separates the two classes would have 100 % sensitivity and specificity / A completely ineffective model would result in an ROC curve that closely follows the 45 ◦ diagonal line and would have an area under the ROC curve of approximately 0.50.) Area under the curve can be used as a quantitative measure of performance
- *When to use*: Helpful tool for choosing a threshold that appropriately maximizes the trade-off between sensitivity and specificity (e.g., Lowering the threshold (aka 50%) can we improve the sensitivity to capture more true positives). Make a quantitative assessment of the model
- *Advantage*: the curve is insensitive to disparities in the class proportions. Metrics that is independent of probability cutoffs
- *Disadvantage*: disadvantage of using the area under the curve to evaluate models is that it obscures information (i.e., the curves cross both AUC can be the same).
- *Description*: AUC-ROC measures the area under the receiver operating characteristic curve, which plots the true positive rate (recall) against the false positive rate at various classification thresholds (10%, 20%… 50% = commonly used).
- *Example*: An AUC-ROC score of 0.85 indicates that the model has an 85% chance of correctly ranking a randomly chosen positive instance higher than a randomly chosen negative instance.
- *Calculation*: AUC-ROC can be calculated using various methods, such as the trapezoidal rule or Mann-Whitney U statistic.
- *Notes*:
  - We can use the partial area under the ROC curve as a technique to summarize these curves that focuses on specific parts of the curve.
  - ROC technique can be extended to fit three or more classes problems

6) **Lift Charts**:

- *Higher Value*: Better performance (Figure)
- *When to use*: To assess the ability of a model to detect events in a data set with two classes and allow us to choose a quasithreshold for a model.
- *Advantage*: Easy connect the model to the buisness value: Using the lift plot, the expected profit can be calculated for each point on the curve to determine if the lift is sufficient to beat the baseline profit
- *Disadvantage*: Bad for comparing different models

- *Description*: The lift chart plots the cumulative gain/lift against the cumulative percentage of samples that have been screened
- *Example*: Figure shows the best and worse case lift curves for a data set with a 50 % event rate. The non-informative model has a curve that is close to the 45 ◦ reference line, meaning that the model has no benefit for ranking samples. The other curve is indicative of a model that can perfectly separate two classes. At the 50 % point on the x-axis, all of the events have been captured by the model.
- *Calculation*: *NA*
- *Notes*:
  - The section of the curve associated with the highest-ranked samples should have an enriched true-positive rate and is likely to be the most important part of the curve.

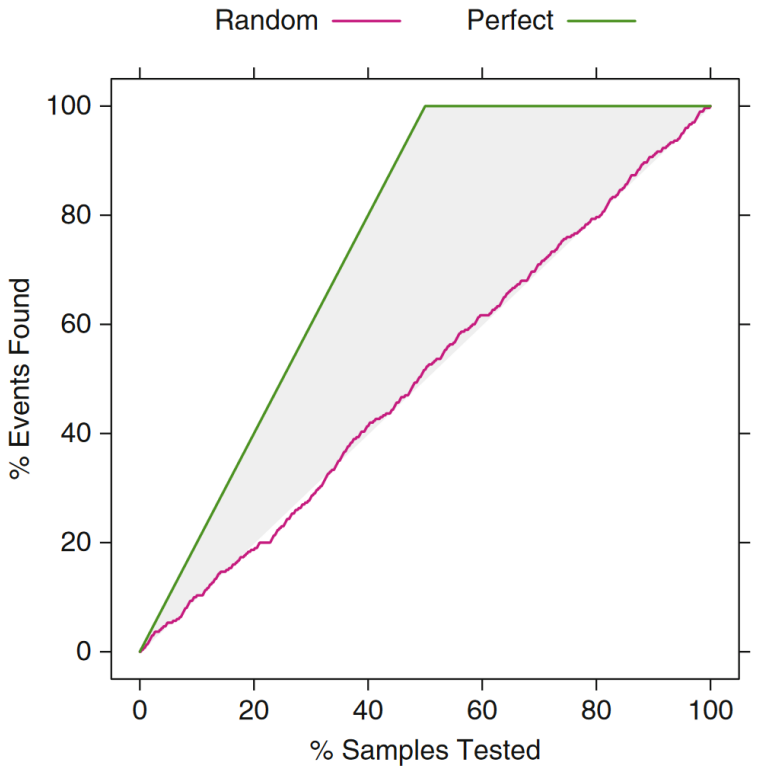

Fig. 11.7: An example lift plot with two models: one that perfectly separates two classes and another that is completely non-informative

Figure 6.5: Lift Charts

NOTES:

- It is important to test whether the estimated class probabilities are reflective of the true underlying probability of the sample (well-calibrated Probabilities) using a *calibration*

plot. This plot shows some measure of the observed probability of an event versus the predicted class probability. One approach for creating this visualization is to score a collection of samples with known outcomes (preferably a test set) using a classification model. The next step is to bin the data into groups based on their class probabilities. For example, a set of bins might be [0, 10 %], (10 %, 20 %], ..., (90 %, 100 %]. For each bin, determine the observed event rate. Suppose that 50 samples fell into the bin for class probabilities less than 10 % and there was a single event. The midpoint of the bin is 5 % and the observed event rate would be 2 %. The calibration plot would display the midpoint of the bin on the x-axis and the observed event rate on the y-axis. If the points fall along a $45\circ$ line, the model has produced well-calibrated probabilities.

- If there are three or more classes, a heat map of the class probabilities can help gauge the confidence in the predictions.
- An approach to improving classification performance is to create an *equivocal* or *indeterminate zone* where the class is not formally predicted when the confidence is not high. (e.g., For a two-class problem that is nearly balanced in the response, the equivocal zone could be defined as 0.50 ± z.Ifz were 0.10, then samples with prediction probabilities between 0.40 and 0.60 would be called "equivocal." In this case, model performance would be calculated excluding the samples in the indeterminate zone.)

6.3.1.3 Non-Accuracy-Based Criteria

When accuracy is not the primary goal for the predictive model and we want to quantify the consequences of correct and incorrect predictions (i.e., the benefits and costs)

Examples:

- Predict investment opportunities that maximize return
- Improve customer satisfaction by market segmentation
- Lower inventory costs by improving product demand forecasts
- Reduce costs associated with fraudulent transactions: For example, in fraud detection, a model might be used to quantify the likelihood that a transaction is fraudulent. Suppose that fraud is the event of interest. Any model predictions of fraud (correct or not) have an associated cost for a more in-depth review of the case. For true positives, there is also a quantifiable benefit to catching bad transactions. Likewise, a false negative results in a loss of income.

1) **profit** = *Cost/Benefit* * TP − *Cost/Benefit* FP − *Cost/Benefit* FN
2) **NEC** (normalized expected cost / classification_cost_penalized) = PCF × (1 − TP)+(1− PCF) × FP (between 0 and 1)

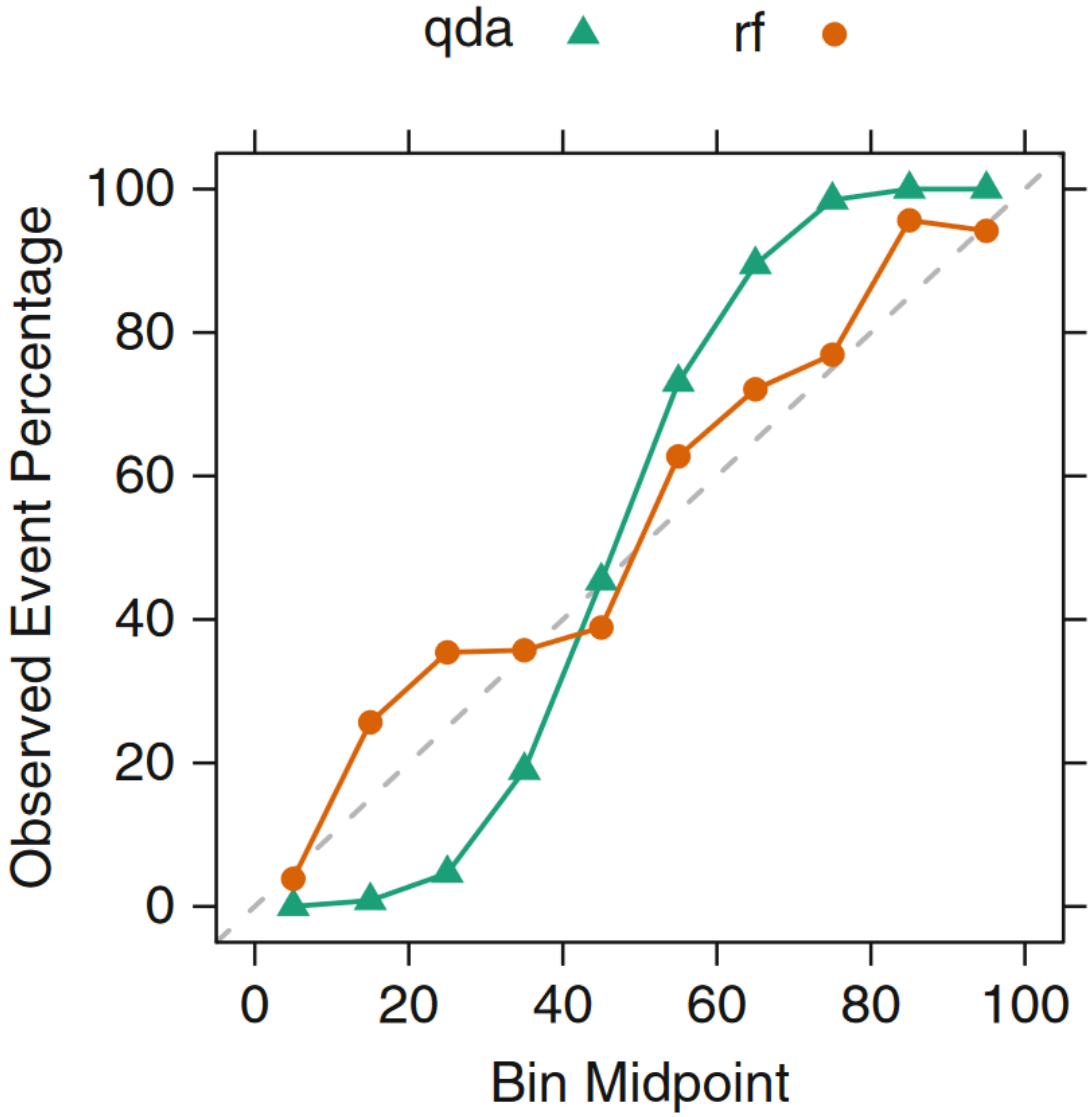

Figure 6.6: A calibration plot of the test set probabilities for random forest and quadratic discriminant analysis models

6.3.2 For models predicting a numeric outcome

1) **RMSE**:

- *Higher Value*: Worse performance
- *When to use*: Commonly used to measure the average magnitude of prediction errors.
- *Advantage*: Penalizes larger errors more heavily, sensitive to outliers and unit is the same as the target variable, making it more interpretable.
- *Disadvantage*: Sensitive to outliers.
- *Description*: The average distance between the observed values and the model predictions.
- *Example*: Continuing with the house price prediction example, an RMSE of 100 means that, on average, the predicted house prices deviate from the actual prices by $100.
- *Calculation*: Squared root of the (sum the residuals (the observed values minus the model predictions) and dividing by the number of samples) For example, if we have actual values [5, 10, 15] and predicted values [6, 12, 10], the MSE would be calculated as ((1^2) + (2^2) + (5^2)) / 3 = 10.

2) **MAE**:

- *Higher Value*: Worse performance
- *When to use*: Suitable when you want to avoid the influence of outliers.
- *Advantage*: Not sensitive to outliers as it uses the absolute error.
- *Disadvantage*: It does not penalize large errors as heavily as RMSE.
- *Description*: The Mean Absolute Error measures the average of the absolute differences between predicted and actual values.
- *Example*: For the house price prediction, an MAE of $50 means that, on average, the predicted house prices deviate from the actual prices by $50.
- *Calculation*: For example, with the same actual and predicted values, the MAE would be calculated as (|1| + |2| + |5|) / 3 = 2.67.

3) **R2**:

- *Higher Value*: Better performance
- *When to use*: Commonly used to measure the average magnitude of prediction errors. It is a measure of correlation, not accuracy. **Bad** for predicting a **number** (accuracy) but **good** for determining the **rank** correlation between the observed and predicted values (e.g., pharmaceutical scientists want to find the compounds predicted to be the most biologically active).
- *Advantage*: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.
- *Disadvantage*: It can be misleading when used with complex models or when the number of predictors is large. It is dependent on the variation in the outcome (e.g., If the range of the houses in the test set was large, say from $60K to $2M, the variance of the sale

price would also be very large. One might view a model with a 90 % R2 positively, but the RMSE may be in the tens of thousands of dollars—poor predictive accuracy for anyone selling a moderately priced property)

- *Description*: The proportion of the information in the data that is explained by the model
- *Example*: An R-squared of 0.75 means that 75% of the variance in the house prices can be explained by the model, and the remaining 25% is due to random variation.
- *Calculation*: Correlation coefficient between the observed and predicted values
- *Note*: By plotting R2 we can see where the model is overpredict (e.g., low values) and underpredict (e.g., higher values). If this happend depending on the context, this systematic bias in the predictions may be acceptable if the model otherwise works well.

4) **R2 adjusted**:

- *Higher Value*: Better performance
- *When to use*: Helpful when you have multiple predictors and want to account for model complexity.
- *Advantage*: It adjusts R-squared for the number of predictors, giving a more reliable assessment of model performance when compared to R-squared.
- *Disadvantage*: It might not penalize overfitting adequately with large numbers of predictors.
- *Description*: R-squared adjusted is similar to R-squared but takes into account the number of predictors in the model. It penalizes models with more predictors if they don't contribute significantly to the variance explained.
- *Example*:
- *Calculation*:

5) **MAPE**:

- *Higher Value*: Worse performance
- *When to use*: Useful when you want to evaluate the performance in percentage terms.
- *Advantage*: Represents the percentage difference between predicted and actual values, making it interpretable and independent of the scale of the data.
- *Disadvantage*: It can be problematic when actual values are close to zero.
- *Description*: The Mean Absolute Percentage Error calculates the mean percentage difference between predicted and actual values.
- *Example*: An MAPE of 10 means that, on average, the predicted house prices deviate from the actual prices by 10%.
- *Calculation*: For example, if we have actual values [100, 50, 75] and predicted values [90, 40, 70], the MAPE would be calculated as (|(100-90)/100| + |(50-40)/50| + |(75-70)/75|) / 3 0.16.

6) **EV**:

- *Higher Value*: Better performance (good value > 0.6)
- *When to use*: Useful to understand how well the model explains the variance in the target variable.
- *Advantage*: Measures the proportion of variance explained by the model, similar to R-squared.
- *Disadvantage*: It might not penalize the model adequately for underfitting or overfitting.
- *Description*: The Explained Variance Score quantifies the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with 1 indicating a perfect fit. For example, an EV of 0.85 means that 85% of the variance is explained by the model.
- *Example*: An EV of 0.9 means that the model explains 90% of the variance in the house prices, leaving 10% unexplained by the model.
- *Calculation*:

7) **MSLE**:

- *Higher Value*: Worse performance
- *When to use*: Suitable when you want to focus on the ratio of errors rather than their absolute differences. It can be useful when predictions are on a large scale.
- *Advantage*: Penalizes underestimation and overestimation proportionally and is less sensitive to large errors.
- *Disadvantage*: The logarithmic transformation can be problematic for data containing zero or negative values.
- *Description*: The Mean Squared Logarithmic Error calculates the mean of the squared logarithmic differences between predicted and actual values.
- *Example*: For the house price prediction, an MSLE of 0.1 means that, on average, the predicted house prices deviate from the actual prices by 10% when measured on a logarithmic scale.
- *Calculation*: For instance, if we have actual values [100, 50, 75] and predicted values [110, 40, 80], the MSLE would be calculated as ((log(110) - log(100))^2 + (log(40) - log(50))^2 + (log(80) - log(75))^2) / 3 0.015.

7 Model Choosing

Once the settings for the tuning parameters have been determined for each model, the question remains: how do we choose between multiple models?

1) Start with several models that are the least interpretable and most flexible, such as boosted trees or support vector machines. Across many problem domains, these models have a high likelihood of producing the empirically optimum results (i.e., most accurate).
2) Investigate simpler models that are less opaque (e.g., not complete black boxes), such as multivariate adaptive regression splines (MARS), partial least squares, generalized additive models, or näıve Bayes models.
3) Consider using the simplest model that reasonably approximates the performance of the more complex methods.

NOTE: A paired t-test can be used to evaluate if the differences between models are statistically significant. It is also recommended to plot confidence intervals that were derived using the bootstrap (Figure) fot two reasons.

- The interval quantifies the variation in the model but is also reflective of the data. For example, smaller test sets or noise (or mislabeling) in the response can lead to wider intervals.
- Facilitate trade-offs between models. If the confidence intervals for two models significantly overlap, this is an indication of (statistical) equivalence between the two and might provide a reason to favor the less complex or more interpretable model.

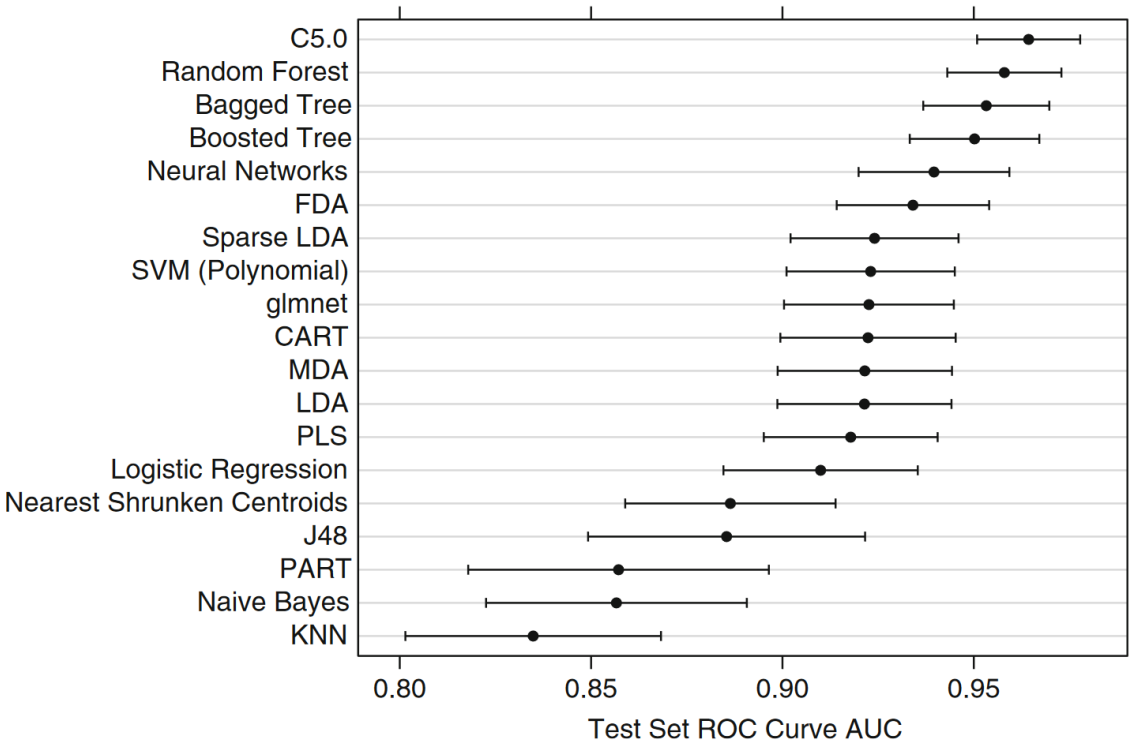

Figure 7.1: A plot of the test set ROC curve AUCs and their associated 95 % confidence intervals

8 Linear Regression and Its Cousins

They all seek to find estimates of the parameters to minimize the sum-of-squared errors

8.1 Ordinary linear regression

Finds parameter estimates that have minimum bias using the NIPALS approach **Advantages**:

- highly interpretable
- enables us to compute standard errors of the coefficients allowing to assess the statistical significance of each predictor

8.2 Partial least squares (PLS)

For regression and classification:

supervised dimension reduction procedure while PCR (PCA + linear regression) is **unsupervised**

USE: when there are correlated predictors and a linear regression-type solution is desired instead of PCA then linear regression (AKA PCR; If, the variability in the predictor space is not related to the variability of the response, then PCR can have difficulty identifying a predictive relationship when one might actually exist).

Efficiently for data sets of small-to moderate size (e.g., $< 2{,}500$ samples and < 30 predictors)

Pre-prossesing:

- centered and scaled predictors.
- Remove predictors with small PLS regression coefficients and small VIP (<1)
- *To include nonlinear relationships add squared or cubic predictors*
- *To include nonlinear relationships splits each predictor into two or more bins for those predictors that are thought to have a nonlinear relationship with the response. Cut points for the bins are selected by the user and are based on either prior knowledge or characteristics of the data. The original predictors that were binned are then excluded from the data set that includes the binned versions of the predictors. (GIFI approach)*

Tuning: Cross-validation was used to determine the optimal number of PLS components to retain that minimize RMSE *# of tuning parameter*: PLS has one tuning parameter: the number of components to retain

For classification: **NOTES**: Produce continuous predictions that do not follow the definition of a probability-the predicted values are not necessarily between 0 and 1 and do not sum to 1. Therefore, a transformation (e.g., *softmax transformation*) must be used to coerce the predictions into "probability-like" values so that they can be interpreted and used for classification.

8.2.1 Algorithmic Variations of PLS

- SIMPLS approach

USE: for data large data sets (e.g., > 2,500 samples and > 30 predictors)

- Rannar et al. (1994) kernel

USE: when there are more predictors than samples.

8.3 Penalized models

For classification and regressions: Finds parameter estimates that have lower variance. We introduce bias to reduce variance and avoid overfitting

USE: When sample size are small

Advantages: reduce variance and increases prediction on the long term

8.3.1 Ridge regression

https://www.youtube.com/watch?v=Q81RR3yKn30

Advantages: better at reducing variance in models that contain usefull variables

8.3.2 Lasso regression

https://www.youtube.com/watch?v=NGf0voTMlcs

Advantages: - better at reducing variance in models that contain useless variables - simplify model

8.3.3 Elastic net

https://www.youtube.com/watch?v=1dKRdX9bfIo

USE: When you don't know if you have useless variables **Advantages**: Best of both ridge and lasso

8.3.4 Logistic Regression

For classification only https://www.youtube.com/watch?v=yIYKR4sgzI8&list=PLblh5JKOoLUKxzEP5HA2d-Li7IJkHfXSe

8.3.5 Linear Discriminant Analysis

For classification only

It is like PCA but it focuses on maximizing the separability among the known categories.

Create an axis (create two axis for three of more categories) that maximizes the distance between the means for the two categories while minimizing the scatter.

As in PCA the first axis created (LDA1) by LDA accounts for the most variation between the categories. LDA2 does the second better job, LDA 3 the third best job etc etc...

NOTE: We can see which variables correlates the most with each LDA

https://www.youtube.com/watch?v=azXCzI57Yfc

8.3.6 Partial Least Squares Discriminant Analysis

For classification only

8.3.7 Nearest Shrunken Centroids

For classification only

9 Nonlinear Models

In these models, the exact form of the nonlinearity does not need to be known explicitly or specified prior to model training.

9.1 Tree-based models

Advantages: - Handles high-dimensional data well. - Robust to outliers and noise. - Provides feature importance measures. - Requires minimal data preprocessing and is relatively easy to implement.

Disadvantages:

- Can be computationally expensive for large datasets.
- May not perform well on imbalanced datasets.
- Lacks interpretability compared to single decision trees.

USE:

- When you have a large dataset with a high number of features.
- When interpretability is not a top priority.
- When you want to build a model that is robust to overfitting.

Tuning parameters:

- The number of trees in the forest. Higher values generally improve performance but increase computational time (Common values are between 50 to 500).
- The maximum depth of each decision tree: Controls the tree's complexity and potential overfitting (Common values are between 5 to 50).
- The minimum number of samples required to split an internal node: Higher values prevent overfitting Common values are between 2 to 20).
- The number of features to consider when looking for the best split: Common values are 'sqrt' (square root of total features) or 'log2'.

For regression:

https://www.youtube.com/watch?v=g9c66TUylZ4

For classification:

https://www.youtube.com/watch?v=_L39rN6gz7Y

9.1.1 Random forests

https://www.youtube.com/watch?v=J4Wdy0Wc_xQ

9.1.2 Boosting trees

9.1.2.1 AdaBoost

Advantages:

- Can achieve high accuracy by combining multiple weak learners.
- Can handle both classification and regression problems.
- Less susceptible to overfitting compared to individual decision trees.
- Can be combined with any base estimator that accepts sample weights.

Disadvantages:

- Sensitive to noisy data and outliers.
- Can be computationally expensive as it requires sequentially training multiple learners.
- May not perform well on highly imbalanced datasets.

USE:

- When you have a moderately sized dataset and you want to improve the accuracy of weak learners.
- When you want to create a powerful ensemble with different weak learners.

Tuning parameters:

- n_estimators: The number of boosting stages (weak learners) to be run. Common values are between 50 to 500.
- learning_rate: The contribution of each weak learner to the final combination. Common values are between 0.01 to 1.0.
- base_estimator: The base estimator used for boosting. Common choices are decision trees with max_depth set or linear models.

https://www.youtube.com/watch?v=LsK-xG1cLYA

9.1.2.2 XgBoost

Advantages:

- Highly efficient and scalable, making it suitable for large datasets.
- Can handle missing data and supports regularization to prevent overfitting.
- Provides built-in cross-validation, early stopping, and feature importance.
- Often performs well even with default hyperparameters.

Disadvantages:

- Can be sensitive to hyperparameter tuning.
- Requires more careful tuning and validation compared to Random Forest and AdaBoost.
- The interpretation of feature importance may not be as straightforward as in Random Forest.

USE:

- When you have a large dataset and computational efficiency is crucial.
- When you need better performance compared to other algorithms on a wide range of problems.
- When you can invest time in tuning hyperparameters.

Tuning parameters:

- n_estimators: The number of boosting rounds. Common values are between 50 to 500.
- learning_rate: The step size shrinkage used to prevent overfitting. Common values are between 0.01 to 0.3.
- max_depth: The maximum depth of each tree. Common values are between 3 to 10.
- subsample: The fraction of samples used for training each tree. Common values are between 0.5 to 1.0.
- colsample_bytree: The fraction of features used for training each tree. Common values are between 0.5 to 1.0.

For regression: https://www.youtube.com/watch?v=3CC4N4z3GJc https://www.youtube.com/watch?v=OtD8wVaFm6E

For classification: https://www.youtube.com/watch?v=jxuNLH5dXCs https://www.youtube.com/watch?v=8b1JEDvenQU

Tuning parameters:

- maximum number of leaves: (generally between 8 and 32)
- learning rate: Value between 0 and 1 (generally 0.1)

9.2 Multivariate Adaptive Regression Splines (MARS)

MARS uses surrogate features (usually a function of only one or two predictors (second degree) at a time which are broken into two groups and models linear relationships between the predictor and the outcome in each group) instead of the original predictors (like pls and neural networks).

NOTES - GCV statistic is use to determine the contribution of each feature to the model.

Tuning parameters:

- the degree of the features that are added to the model (number of interaction; e.g., 0-4)
- the number of retained terms

Advantages:

- the model automatically conducts feature selection
- interpretability is high
- requires very little pre-processing of the data (Correlated predictors do not drastically affect model performance, but they can complicate model interpretation.).

Disadvantages: For MARS models that can include two or more terms at a time, we have observed occasional instabilities in the model predictions where a few sample predictions are wildly inaccurate (perhaps an order of magnitude off of the true value). This problem has not been observed with *additive MARS* models (models with degree of 1).

USE: When there is a clear indication that the relationship between the dependent variable and independent variables is non-linear and interpretability is impportante

9.3 Support Vector Machines (SVM)

USE: When we seek to minimize the effect of outliers

https://www.youtube.com/watch?v=efR1C6CvhmE

NOTE: This principle also apply for regression. However in this case the svm will search for hyperplane that holds the maximum of the observation within the margin (tolerance level)

Tuning parameters: - Kernel: SVM can use different kernel functions to transform the input data into a higher-dimensional space, where it becomes easier to find a separating hyperplane. Common kernel functions include: - Linear Kernel (If regression line is truly linear, the linear kernel function will be a better choice) - Polynomial Kernel (In general, quadratic models have smaller error rates than the linear models) (tuning parameters: degree and scale factor (coef0) and c) - Radial Basis Function (RBF) Kernel (radial basis function has been shown to be

very effective overall and easier to tune than polynomial one less tuning parameter) (tuning parameters: (sigma) that controls the scale and c) - hyperbolic tangent

- Threshold (epsilon) (called margin in tidymodels?) (If the threshold is set to a relatively large value, then the outliers are the only points that define the regression line) (e.g., = 0.01; the cost parameter provides more flexibility for tuning the model. So it is suggested to fix a value for and tune over the other kernel parameters)
- C parameter (Cost; e.g., values between 0.25 and 2048):
  - *For classification*: It controls the trade-off between maximizing the margin and minimizing the classification error. A smaller C value creates a wider margin but may allow some misclassifications, while a larger C value creates a narrower margin but may result in fewer misclassifications on the training set.
  - *For regression*: The cost parameter is the main tool for adjusting the complexity of the model. When the cost is large, the model becomes very flexible since the effect of errors is amplified. When the cost is small, the model will "stiffen" and become less likely to over-fit (but more likely to underfit)

Pre-processing: Center and scale the predictors prior to building an SVM model since the predictors enter into the model as the sum of cross products, differences in the predictor scales can affect the model.

9.4 K-Nearest Neighbors (KNN)

For classification:

For regression:

Tuning parameters: K number of neighbors.

Advantages: The KNN method can have poor predictive performance when local predictor structure is not relevant to the response.

Pre-processing: - Remove irrelevant, noise-laden predictors is a key pre-processing step for KNN, since these can cause similar samples to be driven away from each other in the predictor space

NOTE: to enhance KNN predictability weight the neighbors' contribution to the prediction of a new sample based on their distance to the new sample.

9.5 Neural Networks

Neural Networks uses surrogate features instead of the original predictors (like pls and MARS)

For classification: **NOTES**: Produce continuous predictions that do not follow the definition of a probability-the predicted values are not necessarily between 0 and 1 and do not sum to 1. Therefore, a transformation (e.g., *softmax transformation*) must be used to coerce the predictions into "probability-like" values so that they can be interpreted and used for classification.

For regression: *To be continued*

9.6 Nonlinear Discriminant Analysis

For classification only:

9.7 Flexible Discriminant Analysis

For classification only:

9.8 Naive Bayes

For classification only:

10 Tree based models

11 Treebased classification models

Characteristics: - resistant to outliers

12 Treebased regression models

13 Remedies for suboptimal data

13.1 Class Imbalance

An imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes.

- Online advertising: Ad clicked or not (2.4%)
- Pharmaceutical research: Molecules with activity (vwry few) or not
- Insurance claims: Fraud (only 22%) or not fraud
- Spam detection: Spam or not spam
- Selling buisness: Buy (6%) or not buy

13.1.1 The Effect of Class Imbalance

- The models achieve good specificity (since almost every customer is predicted no insurance) but have poor sensitivity (Figure).

Table 16.1: Results for three predictive models using

Model	Accuracy	Kappa	Sensitivity	Spec
Random forest	93.5	0.091	6.78	
FDA (MARS)	93.8	0.024	1.69	
Logistic regression	93.9	0.027	1.69	

- The imbalance also had a severe effect on the predicted class probabilities. (e.g., In the random forest model, for example, 82 % of the customers have a predicted probability of having insurance of 10 % or less. This highly left-skewed predicted probability distribution also occurs for the other two models. This means that the models are not very confident in predicting that most customers have insurance; they tend to assign low probabilities of having insurance to a significant portion of the customers.)
- Imbalance cause that lift charts and ROC curves have similar patterns

13.1.2 Strategies for overcoming class imbalances

13.1.3 Hyperparameters selection

- **Model tuning strategy**: tune the model to maximize the accuracy or sensitivity of the minority class(es)

13.1.4 Post-processing techniques (use model outputs)

- **Alternate probability Cutoffs** to improve the prediction accuracy of the minority class samples (i.e., post-processing the model predictions to redefine the class predictions). The most straightforward approach is to use the *ROC curve* since it calculates the sensitivity and specificity across a continuum of cutoffs. Using this curve, an appropriate balance between sensitivity and specificity can be determined.

  – Several techniques exist for determining a new cutoff:

  1) First, if there is a particular target that must be met for the sensitivity or specificity, this point can be found on the ROC curve and the corresponding cutoff can be determined.
  2) Another approach is to find the point on the ROC curve that is closest (i.e., the shortest distance) to the perfect model (with 100 % sensitivity and 100 % specificity), which is associated with the upper left corner of the plot. In Figure, a cutoff value of 0.064 would be the closest to the perfect model.
  3) The cutoff associated with the largest value of the Youden index (measures the proportion of correctly predicted samples for both the event and nonevent groups / can be computed for each cutoff that is used to create the ROC curve): show superior performance relative to the default 50 % value. For the random forest ROC curve, the cutoff that maximizes the Youden index (0.021) is similar to the point closest to the optimal model.

NOTE: In our analysis, the alternate cutoff for the model was not derived from the training or test sets. It is important, especially for small samples sizes, to use an independent data (small evaluation set used for developing post-processing techniques ~ 10% training set used to tune model) set to derive the cutoff. If the training set predictions are used, there is likely a large optimistic bias in the class probabilities that will lead to inaccurate assessments of the sensitivity and specificity. If the test set is used, it is no longer an unbiased source to judge model performance.

- **Adjusting Prior Probabilities**: For models that use prior probabilities naıve Bayes and discriminant analysis classifiers. Unless specified manually, these models typically derive the value of the priors from the training data. Weiss and Provost (2001a) suggest that priors that reflect the natural class imbalance will materially bias predictions to the

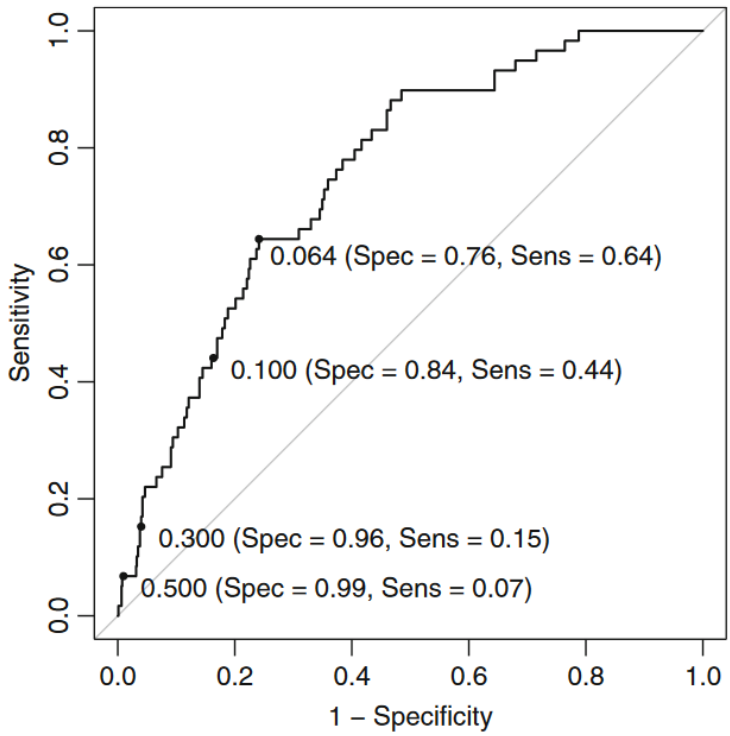

Fig. 16.2: The random forest ROC curve for predicting the classes using the evaluation set. The number on the left represents the probability cutoff, and the numbers in the parentheses are the specificity and sensitivity, respectively. Several possible probability cutoffs are used, including the threshold geometrically closest to the perfect model (0.064)

Figure 13.1: ROC: The predicted sensitivity for the new cutoff of 0.064 is 64.4 %, which is a significant improvement over the value generated by the default cutoff. The consequence of the new cutoff is that the specificity is estimated to drop from 99 % to 75.9 %.

majority class. **Using more balanced priors or a balanced training set may help deal with a class imbalance.** (e.g., when classes are 6 % and 94 % for the insured and uninsured it is better to use 60 % for the insured and 40 % for the uninsured). This strategy did not change the model (same ROC) but allows for different trade-offs between sensitivity and specificity.

13.1.5 Alter training data prior to model training

- **Adjust Sampling Methods**: Non of them is a clear winner, it depends of the case study

  1) *priori sampling approach*: select a training set sample to have roughly equal event rates during the initial data collection. However, the test set should be sampled to be more consistent with the state of nature and should reflect the imbalance so that honest estimates of future performance can be computed.
  2) *post hoc sampling approach*:

  - up-sampling:
    - adding random samples with replacement from minority classes)
  - down-sampling:
    - randomly sample the majority classes so that all classes have approximately the same size
    - bootstrap sample across all cases such that the classes are balanced in the bootstrap set (advatange of bootstrap is that we can obtain the estimate of variation about the down-sampling)
  - SMOTE: adds new/synthetic samples to the minority class and down-sample cases from the majority class via random sampling

  NOTE: Adjusting sampling can bias model performance (e.g., up-sampling: the same sample can be use to predict and tune model)

13.1.6 Alter model training process (model parameters are being modified)

- **Unequal Case Weights**: Many of the predictive models for classification (boosting trees which apply different case weights at each iteration) have the ability to use case weights where each individual data point can be given more emphasis in the model training phase. Increase the weights for the samples in the minority classes
- **Cost-Sensitive Training**: Some models (SVM, CART trees, C5.0 trees) can alternatively optimize a cost or loss function that differentially weights specific types of errors of specific classes (it can cause that class probabilities cannot be generated and ROC cannot be use). For example, it may be appropriate to believe that misclassifying true

events (false negatives) is X times as costly as incorrectly predicting nonevents (false positives). Correctly classify class A is more importante than correctly classify class B

13.2 Big sample size

An increase in the number of samples can have less positive consequences:

- Computational burdens as the number of samples (and predictors) grows
- There are diminishing returns on adding more of the same data from the same population. Since models stabilize with a sufficiently large number of samples, garnering more samples is less likely to change the model fit.

14 Measuring Predictor Importance

Measurement of predictor relevance is derived by permuting each predictor individually and assessing the loss in performance when the effect of the predictor is negated. Useful for guiding the user to focus more closely on specific predictors via visualizations and other means.

Many predictive models have built-in or intrinsic measurements of predictor importance.

- MARS
- Tree-based models

When models have not built-in measurements of predictor importance we can apply some other techniques

14.1 For Numeric Outcomes

- LOESS
- t-statistic
- ANOVA
- Relief

14.2 For Categorical Outcomes

- area under the ROC curve
- t-statistics
- MIC
- Relief

15 Feature Selection

15.1 Unsupervised methods

When the outcome is ignored during the elimination of predictors, the technique is unsupervised.

- removing predictors that have high correlations with other predictors
- removing near-zero variance predictors

15.2 Supervised methods

When, the outcome is typically used to quantify the importance of the predictors. Predictors are specifically selected for the purpose of increasing accuracy or to find a subset of predictors to reduce the complexity of the model

15.3 Consequences of Using Non-informative Predictors

The presence of non-informative variables can add **uncertainty/noise** to the predictions and reduce the overall effectiveness of the model (linear regression, partial least squares, neural networks, svm). Regression trees, MARS models and Random forests are not affected or very slightly in the case of random forests

15.4 Approaches for Reducing the Number of Predictors

- Wrapper methods evaluate multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance. In essence, wrapper methods are search algorithms that treat the predictors as the inputs and utilize model performance as the output to be optimized.
  - **Forward, Backward, and Stepwise Selection**: Add, remove predictors or both to find the model that results in the smallest model RMSE/AIC (Used in linear regressions)

- **Correlation-based feature selection**: find the best subset of predictors that have strong correlations with the outcome but weak between-predictor correlations
- **Simulated annealing**: Starting from an initial solution, the method iteratively explores neighboring solutions with the ability to accept worse solutions initially, gradually decreasing this acceptance as the process continues. This balance between exploration and exploitation helps the algorithm escape local optima and search for global optima in the solution space.
- **Genetic Algorithms**: GAs start with a population of potential solutions encoded as "genetic" representations. Through multiple generations, solutions are selected based on their fitness, which measures how well they solve the problem. These selected solutions then undergo crossover (combination of genetic material) and mutation (random changes) to produce a new population. Over successive generations, the algorithm converges towards better solutions as traits from successful solutions propagate and refine in the population.

Advantages: **Disadvantages**:

- many models are evaluated (which may also require parameter tuning) and thus an increase in computation time
- increased risk of over-fitting

- Filter methods evaluate the relevance of the predictors outside of the predictive models and subsequently model only the predictors that pass some criterion. For example, for classification problems, each predictor could be individually evaluated to check if there is a plausible relationship between it and the observed classes. Only predictors with important relationships would then be included in a classification model.

  Advantages: more computationally efficient **Disadvantages**:

  - the selection criterion is not directly related to the effectiveness of the model
  - methods evaluate each predictor separately, and, consequently, redundant (i.e., highly-correlated) predictors may be selected and important interactions between variables will not be able to be quantified
  - subjective nature to the procedure. Most scoring methods have no obvious cut point to declare which predictors are important enough to go into the model (In practice, finding an appropriate value for the confidence value may require several evaluations until acceptable performance is achieved.)

Note: When using other search procedures or filters for reducing the number of predictors, there is still a risk. The following situations increase the likelihood of selection bias:

- The data set is small.
- The number of predictors is large (since the probability of a non-informative predictor being falsely declared to be important increases).

- The predictive model is powerful (e.g., black-box models), which is more likely to over-fit the data.
- No independent test set is available

When the data set is large, it is recommended separate data sets for selecting features, tuning models, and validating the final model (and feature set). When training sets are small, proper resampling is critical. When the amount of data is not too small (333 obs), it is recommended setting aside a small test set to double check that no gross errors have been committed.

References

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-6849-3.

"What Is the Difference Between Correlation and Cointegration? Is Cointegration a Good Measure of Risk?" n.d. *Quora*. https://www.quora.com/What-is-the-difference-between-correlation-and-cointegration-Is-cointegration-a-good-measure-of-risk. Accessed July 17, 2023.