March Modeling

Quinn Fargen
Quinn.Fargen@jacks.sdstate.edu
Advised by Dr. Thomas Brandenburger
South Dakota State University Department of Mathematics and Statistics

Decision Tree

- Two types of trees: Regression and Classification.
- Branches on a decision tree are created by partitions.
- X_1, X_2, …, X_p are the p possible variables to partition at.
- Two regions, R_j and R_{j+1} are created from the partition.

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- RSS is the Residual Sum of Squares.
- Pick a split point at t of the selected variable to minimize RSS.
- Trees recursively creates regions within regions

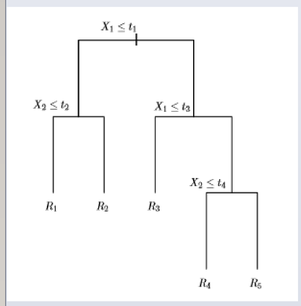

Fig 1. Decision tree example with five nodes and four partitions.

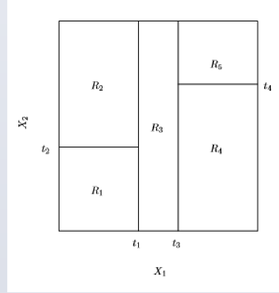

Fig 2. Regions represent the nodes of the corresponding decision tree in Fig 1.

Pruning

- If the tree was allowed, it would create a node for every observation in the dataset.
- Pruning the tree prevents over fitting the dataset.
- Cost-complexity is a common method for regression trees.

$$\sum_{j=1}^{|T|} \sum_{i: x_i \epsilon R_j} (y_i - \hat{y}_{R_j})^2 + \alpha|T|$$

- Error terms for pruning of classification trees are shown in the next column.

Random Forest

- RSS cannot be used for classification trees. Here are three alternatives.

$$\hat{p}_{jk} = \frac{1}{N_j} \sum_{x_i \in R_j} I(y_i = k)$$

- $\hat{p}_{jk}$ is the proportion of observation that are of class k in R_j.
- **Misclassification error:** $= 1 - \hat{p}_{j\,mode(k)} = \frac{1}{N_j} \sum_{x_i \in R_j} I(y_i \neq \max(\hat{p}_{jk}))$
- **Gini Index:** $= \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$
- **Cross-Entropy** $= -\sum_{k=1}^{K} \hat{p}_{jk} \log \hat{p}_{jk}$
- If there are only two classes (Win or Loss), below are the three possible values for these error terms:
- Let p_{j0} = proportion of the first class in R_j. Then p_{j1}= 1- p_{j0}.

Misclassification Error :

$$M = 1 - \hat{p}_{j\,mode(1,2)}$$
$$M(p_{j0}) = 1 - max(p_{j0}, (1 - p_{j0}))$$

Cross-Entropy :

$$C = -(p_{j0}\log(p_{j0}) + p_{j1}\log(p_{j1}))$$
$$C(p_{j0}) = -(p_{j0}\log(p_{j0}) + (1 - p_{j0}))\log(1 - p_{j0}))$$

Gini Index :

$$G = (p_{j0}(1 - p_{j0})) + (p_{j1}(1 - p_{j1}))$$
$$G(p_{j0}) = (p_{j0}(1 - p_{j0})) + (1 - p_{j0})(1 - (1 - p_{j0}))$$
$$= (p_{j0}(1 - p_{j0})) + (p_{j0}(1 - p_{j0}))$$
$$= 2(p_{j0}(1 - p_{j0}))$$

Critical Points :

- Gini Index and Cross-Entropy are both differentiable.
- Each function has a critical point at $\frac{1}{2}$, which are maximums.

- A bootstrap method is used to create training data.
- Each bootstrapped dataset has a random forest built on it.
- B trees are built with m out of the p variables being randomly selected at each new partition. Usually $m = \sqrt{p}$.
- Random forest for regression:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

- Random forest for classification:

$$\hat{C}_{rf}^{B}(x) = mode\left\{\hat{C}_b(x)\right\}_1^B$$

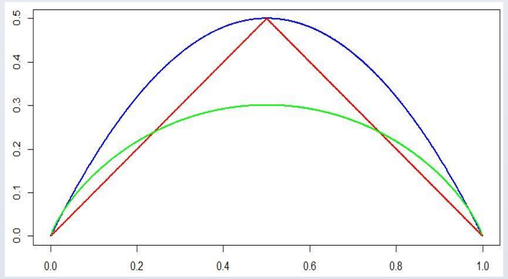

Fig 3. Error curves for Misclassification (Red), Gini Index(Blue), & Cross-Entropy(Green).

March Madness

- 64 teams, 4 regions, 6 rounds, 63 games, 1 winner.
- 9.2 Quintillion ways to fill out each bracket.
- Menzel's ranking from most unlikely of brackets for the last 17 years.
- Menzel based the rankings off of historic outcomes of specific seed matchups.

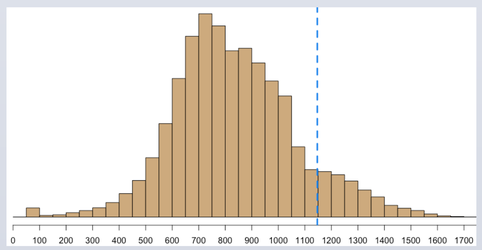

Fig 4. Distribution of 3-million ESPN bracket scores for 2015. Mean of 840 and a SD of 234. (Blue 2015 RF)

1st	2013
2nd	2010
3rd	2012
4th	2014
5th	2011
6th	2006
7th	2000
8th	2001
9th	2016
10th	2002
11th	2005
12th	2015
13th	2004
14th	2003
15th	2008
16th	2009
17th	2007

Bracket Scoring

Round of 64:	10 pts
Round of 32:	20 pts
Sweet 16:	40 pts
Elite 8:	80 pts
Final 4:	160 pts
Championship:	320 pts

Model Variables

- Rank: Position of all NCAA teams
- SOS: Strength of Schedule
- Seed: 1-16 place in bracket region
- OSRS: Offensive simple rating system
- DSRS: Defensive simple rating system
- The rest of the variables are averages across the regular season.

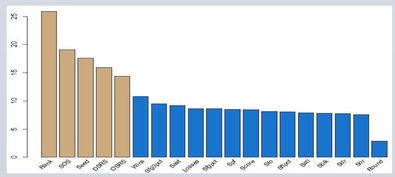

Fig 5. Plot of Importance of variables in Random Forest model. (Gold variables for 5 Variables)

Rank	SOS	Seed	OSRS	DSRS	Wins	Losses	Score	Sfgpct	Sfg3pct	Sftpct	Sor	Sdr	Sast	Sto	Sstl	Sblk	Spf	Round

Bracket Models

- Four different models were used:
- Random Forest & Decision Tree as explained.
- Bagging: Reduced variance by building trees on bootstrapped training datasets.
- Boosting: Reduces variance and bias by building trees off of previous trees created. Can overfit.
- Two sets of variables were used for all four models.
- Models built on 60% tourney games from 2003-2013.
- Models scored on the 2014-2016 tournaments.

Results

19 Variables	2014	2015	2016	Averages (SD)
Random Forest (.265)	811	1147	1296	1085 (212)
Decision Tree (.335)	906	1026	1093	1008 (211)
Bagging (.269)	807	915	1311	1011 (221)
Boosting (.262)	792	984	1315	1030 (217)
Averages (M error)	829 (126)	1018 (142)	1254 (123)	1040 (215)

5 Variables	2014	2015	2016	Averages (SD)
Random Forest (.304)	769	1219	1313	1100 (240)
Decision Tree (.315)	935	1061	1164	1054 (210)
Bagging (.314)	736	1222	1294	1084 (250)
Boosting (.260)	900	916	1263	1026 (169)
Averages (M error)	835 (122)	1104 (169)	1259 (106)	1058 (221)

Conclusions

- Random Forest model performed the best with the 19 and 5 variables. All 5 variable models had higher scores.
- Bagging and Boosting were better than Decision Tree, but not better than Random Forest.
- No perfect brackets are going to come out of these models, but brackets scores show to be above average.