



Spears School
OF BUSINESS

SAS Analytics Day

Predictive Modeling of Titanic Survivors: a Learning Competition

Linda Schumacher

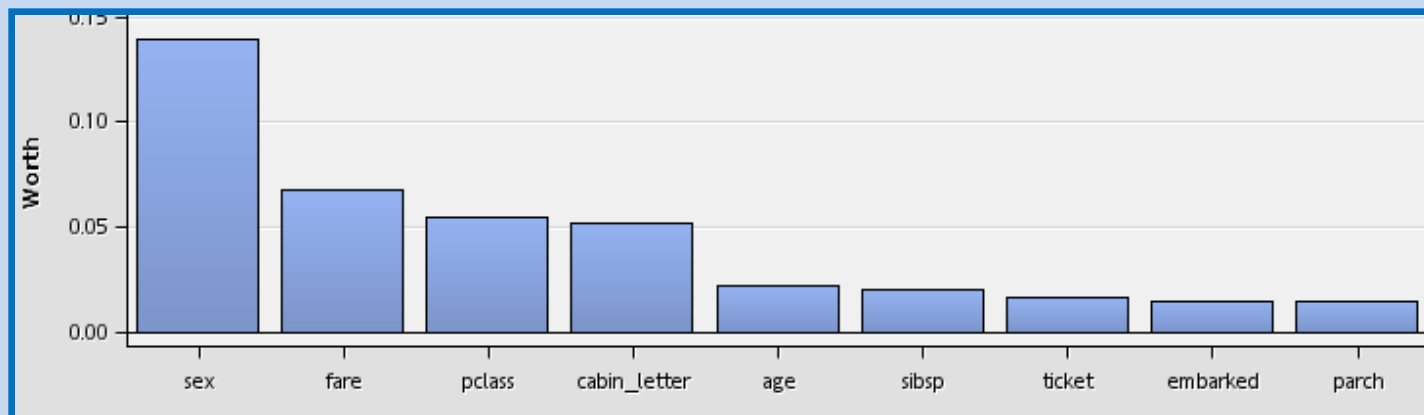
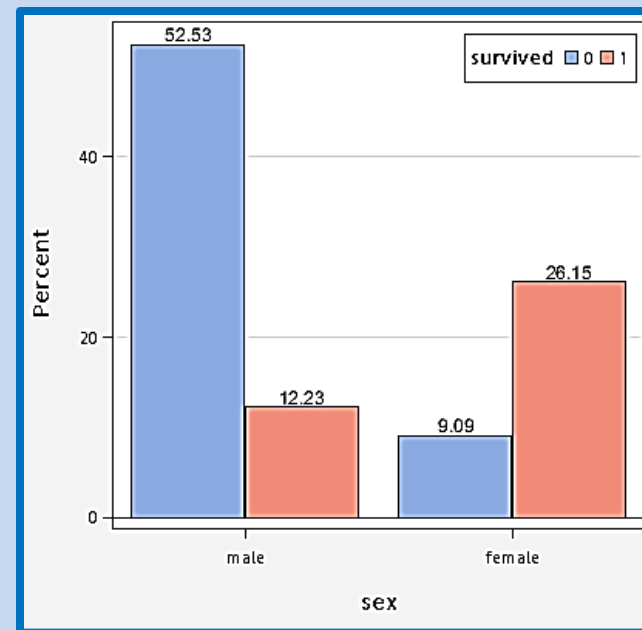


Problem Introduction

- On April 15, 1912, the RMS Titanic sank resulting in the loss of 1502 out of 2224 passengers and crew.
- Predicting the survivors based on demographic variables is a predictive modeling classification problem.
- kaggle.com hosts a Titanic “Getting Starting” public competition with model scoring based on the accuracy fit statistic.
- Predictive Modeling was performed using SAS Enterprise Miner™ 12.1

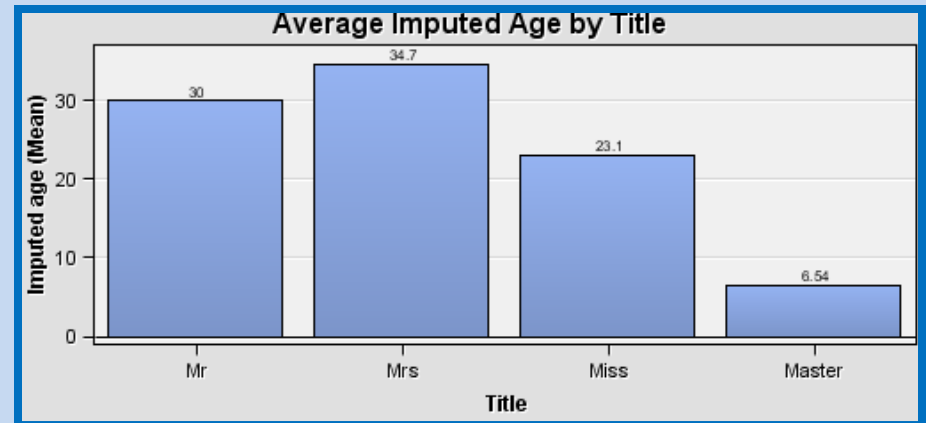
Data Exploration

- Target variable is binary: survived.
- Input variables: name, age, fare, sex, embarked, cabin, ticket, pseudo class, #siblings/spouse, #parents/children.
- Sex, social standing related variables (fare, pclass, cabin) and age have highest worth.



Date Preparation

- Impute Node
- Age has 20% missing values and is imputed with a tree method using a computed variable, title, #siblings/spouse, and #parents/children. Title is extracted from the name.
- Transform Node
- Fare is right skewed with a large range. The log transformation improved the skew and kurtosis.



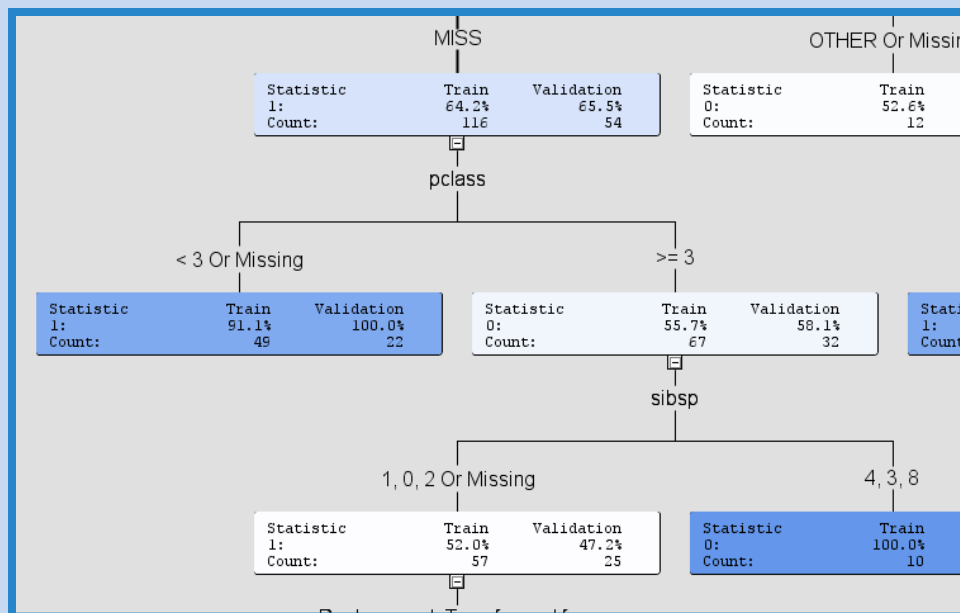
	Range	Skew	Kurtosis
fare	0-512.13	4.79	33.40
log(fare+1)	0 - 6.24	0.42	0.85

Modeling in SAS Enterprise Miner™

- Data Partitioning Node
 - Because of the limited number of cases, 890 passengers, data was partitioned into 70% training and 30% validation.
 - A separate test set of 418 passengers was used for scoring.
- Modeling Nodes
 - Decision Trees, Gradient Boosting Machine, Logistic Regression, Neural Network, Rule Induction.
- Modeling objective is highest accuracy on scored test data.

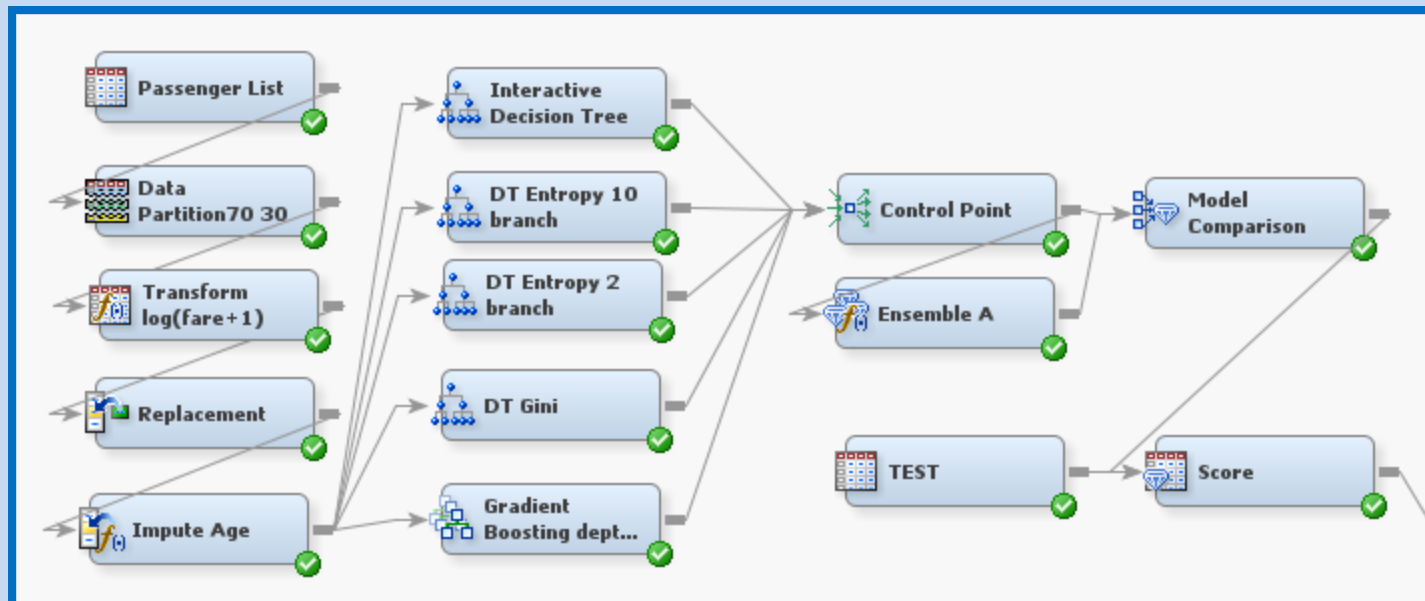
Modeling Variable Importance

- Title is the most important variable identified by trees, GBM, and logistic regression.
- Models next selected variables pclass, #siblings/spouse, ticket, fare, and age.
- Splits on #parents/children, embarked, and cabin rarely occurred.



Modeling Diagram Ensemble A

- Autonomous Decision trees were built using split criteria entropy or Gini. Trees were optimized using the Average Square Error assessment on validation data. An interactive tree was also created.
- A Gradient Boosting Machine with a maximum depth of 10 and a maximum of 2 surrogate rules was created.

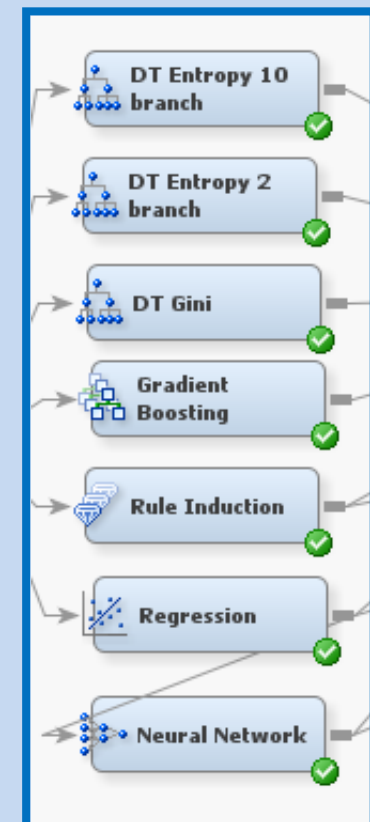


Model Comparison Fit Statistics

- The SAS EM nodes all performed well. The ensembles have the best accuracy fit statistics on scored test data.

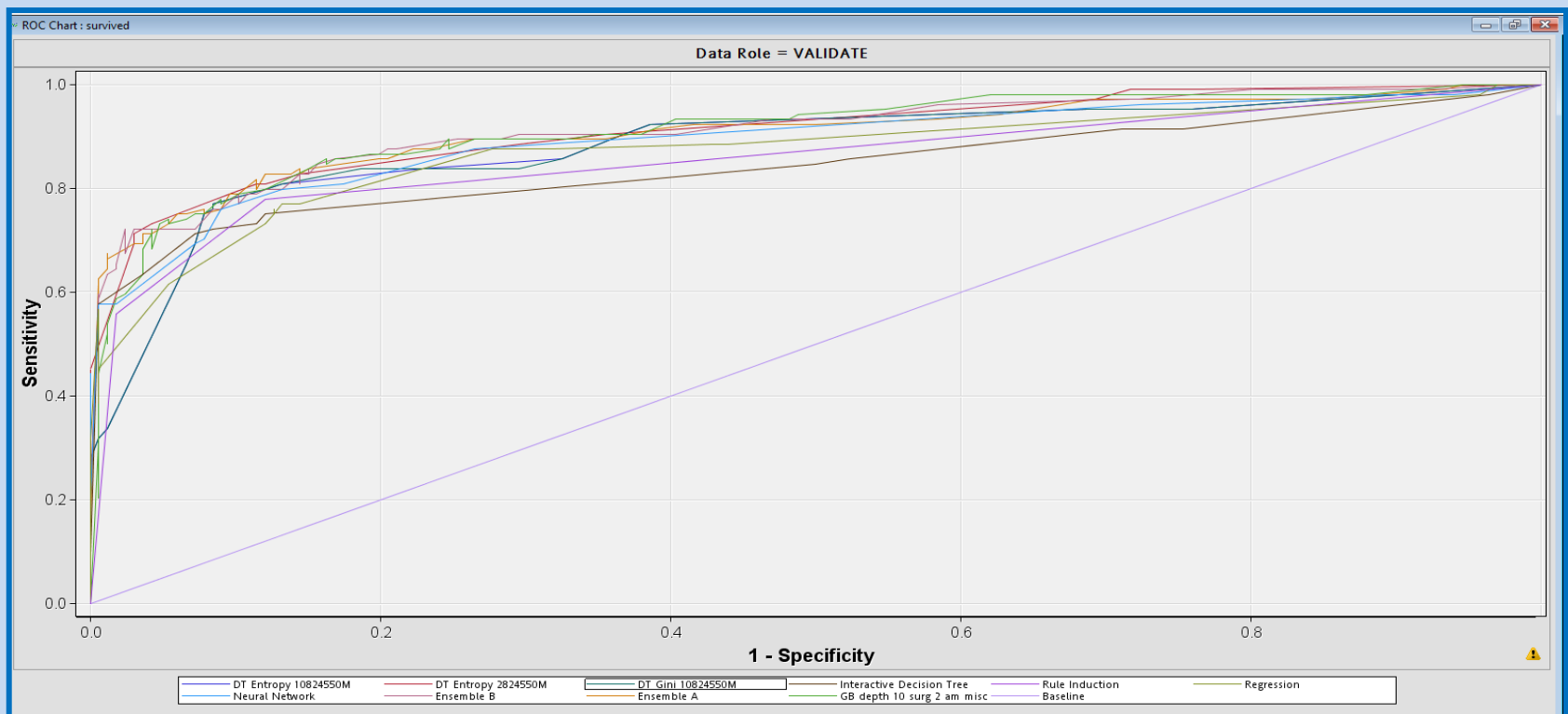
Model	Valid ROC	Valid ASE	Valid MISC	Test Accuracy Score
Ensemble A	0.907	0.104	0.144	0.81340
Ensemble B	0.912	0.105	0.144	0.80816
DT entropy	0.910	0.102	0.129	0.80383
GBM	0.910	0.111	0.133	0.78947
Regression	0.870	0.125	0.178	0.79904
Neural Net	0.892	0.114	0.163	0.76555
Rule Induct	0.856	0.121	0.159	0.79426
DT interact	0.842	0.128	0.156	0.77990

Ensemble B



Model Comparison ROC Chart

- Validation ROC index: 0.842 - 0.912 Test accuracy: 0.766 - 0.813
- An autonomous tree using the entropy split criterion had the highest ROC index. Ensemble A had the highest test accuracy.



Modeling Results

- Title was initially created to impute age but became the most important variable for predictions. Title encapsulates age, gender, marital status, and some professions.
- In general, models predicted survivors with titles: Miss, Mrs., Master
 - Females traveling in 1st & 2nd class and pockets within 3rd class.
 - Males 12 and younger with less than 3 siblings.
- Mr., Rev., and Others were not predicted to survive.
 - Males 13 and older. Note that although 72 men were survivors in training & validation data, the models did not identify any groups of men as survivors.
 - No Reverends in training or validation data survived

Conclusion

- Decision Trees are good predictive models for the Titanic disaster survival because:
 - Variables are related or redundant.
 - Most variables are ordinal or nominal.
 - Trees make no assumptions on the distributions of variables.
 - Interactions between variables are present.
 - Interpretation of tree models is straightforward.
- Ensemble nodes improve on the predictions of component models by forming consensus predictions.

For more details contact presenter at:

Linda Schumacher

Phone: 9198481374

Email: linda.schumacher@okstate.edu

Faculty Advisor: Dr. Goutam Chakraborty

Competition Website: <http://www.kaggle.com/c/titanic-gettingStarted>

Acknowledgements: The author thanks Austen Head, Rick Pack, and Brian Fannin for their valuable comments and suggestions.

