EverBank Data Mining Contest

2013 EverBank Cup

EverBank®

ABOUT CONTEST

TASKS

Predict customer response to mortgage refinance offer. Specifically, students were asked to predict likelihood of customer:

Target1: Filing mortgage refinance application

Target2: Accepting a loan offer from the bank.

DATASETS

- ❑ Training data: ~83k records, including ~27k unique customers and six marketing campaigns.
- ❑ Test data: ~9k customers in a new marketing campaign.

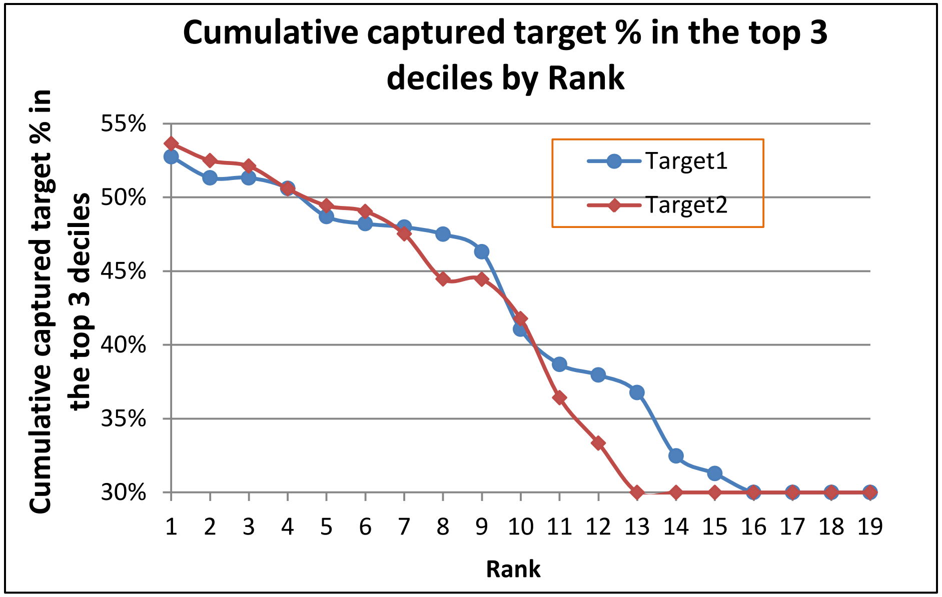
Difficulties: imbalanced data, model stability

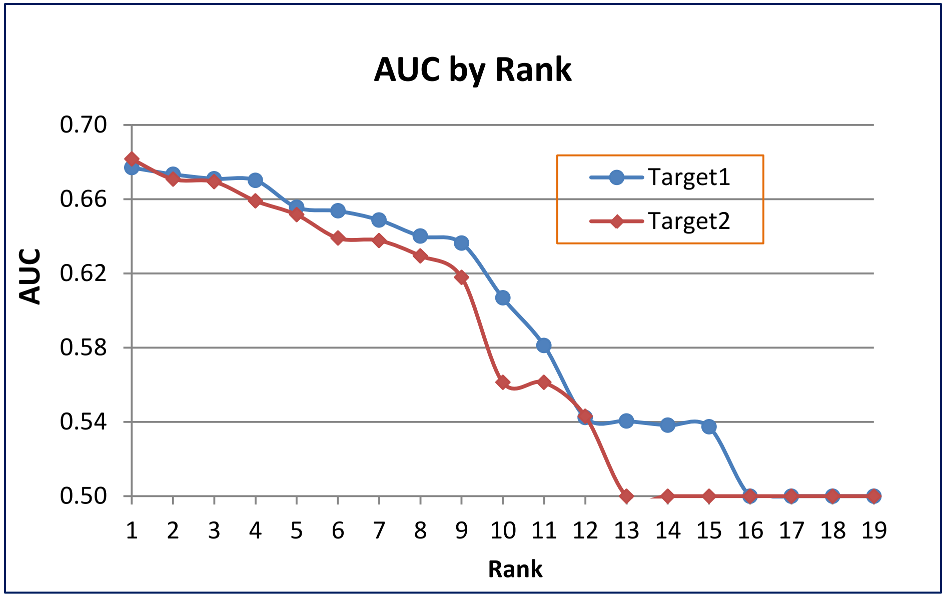
JUDGING RULES

- ❑ 40% - The contents in the project report.
- ❑ 60% - Model accuracy, based on the following:
  - o AUC
  - o Cumulative captured target percent in the top three deciles

RESULTS

- ❑ 33 students registered for the contest.
- ❑ 27 students submitted results.





FINALISTS

Jianbin Zhu

Stephen Jones

Ryan Dagen

2013 EverBank Cup

Winner Presentation

FORECASTING CUSTOMER’S LIKELIHOOD OF LOAN APPLICATION

Jianbin Zhu
Statistics Department
University of Central Florida

BANKING | LENDING | INVESTING

EverBank®

OUTLINE

- Introduction
- Data Analysis Process
- Modeling Approach
- Results and Conclusions

INTRODUCTION

- **Two datasets**

  Training dataset: 83,108 observations and 125 variables (two targets)

  Scoring dataset: 9,138 observations and 123 variables

  The segmentation information based on age, income, education, occupation, marriage status, housing status.

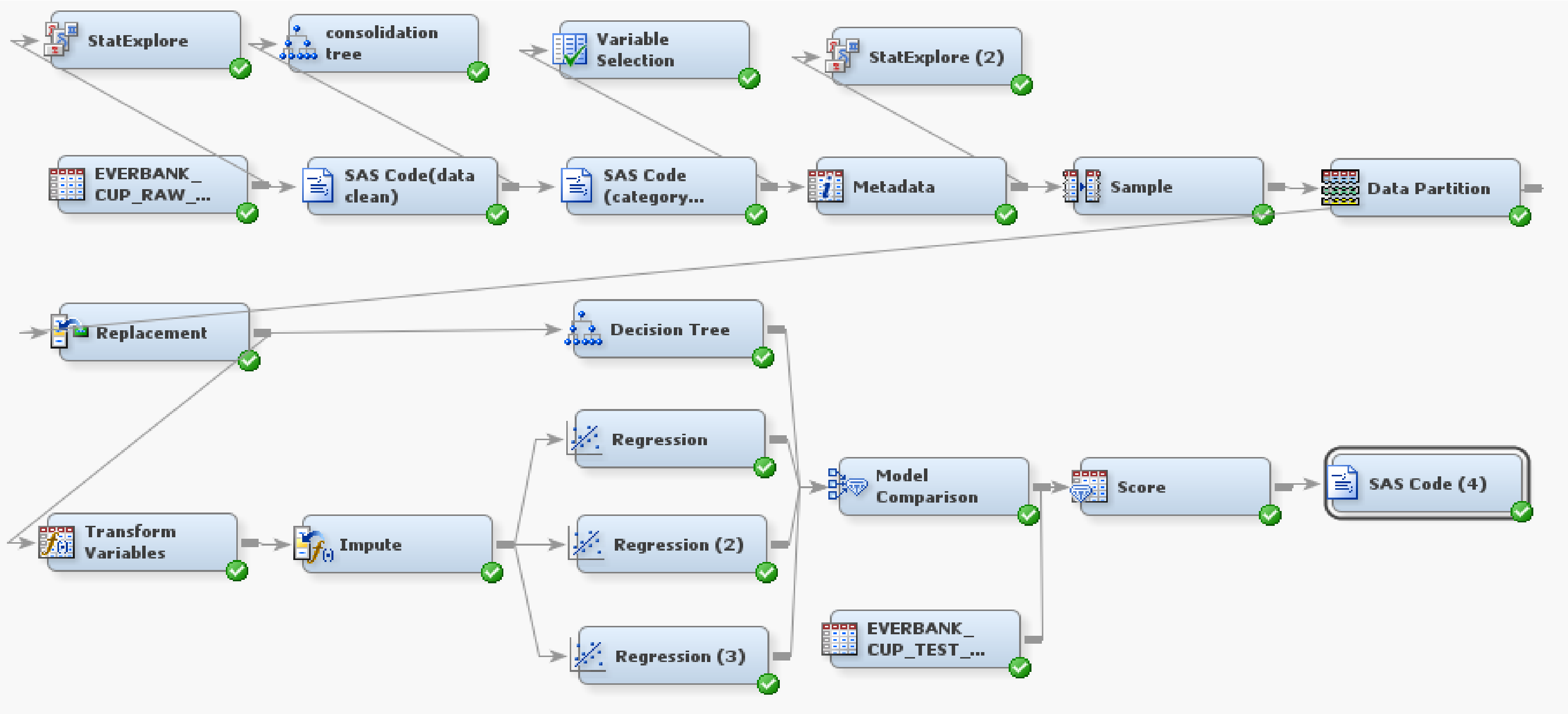
  The custom profiles with credit card, loan situation, interesting rate, month payment and so on.

- **Objective**

  To build a predictive model for each target that will accurately forecast customer's likelihood to respond or get the loan after the next Marketing campaign.

- **Methods**

Data mining techniques and predictive modeling with SAS Enterprise Miner



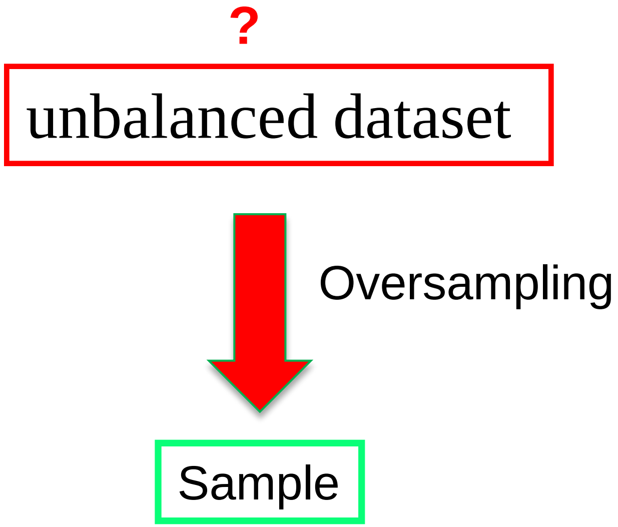
DATA ANALYSIS PROCESS

- **Data Exploration**

a) Targets:

TARGET1	Frequency	Percent
0	76099	91.57
1	7009	8.43

TARGET2	Frequency	Percent
0	79279	95.39
1	3829	4.61



?

unbalanced dataset

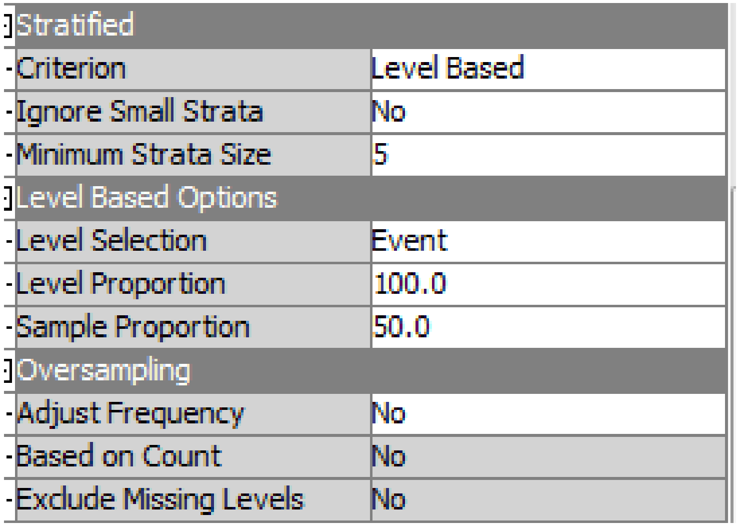
Oversampling

Sample

•Oversampling

- Stratify oversampling method with level based criterion 50/50

Target: Stratification role



Stratified	
Criterion	Level Based
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No

The event of each target is over sample with 50% which is the same as non-event.

b) Categorical variables with high levels

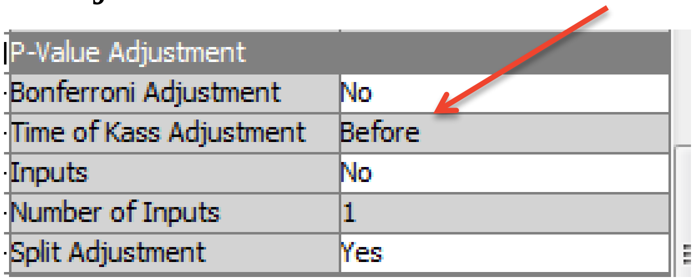
Categorical Consolidation Tree

- Combine the levels that have the same effects on the target.

Field64, 77, 36, 58 and so on

- Decision tree node with a target and a categorical variable

Bonferroni Adjustment is set to NO



P-Value Adjustment	
Bonferroni Adjustment	No
Time of Kass Adjustment	Before
Inputs	No
Number of Inputs	1
Split Adjustment	Yes

- Tree Node in English Rules

convert to SAS Codes

- Repeat for each categorical variable

c) Time variables

Field 2 and Field 32:

Character , Input and Nominal ⇒ Numerical, TimeID and Interval

Field88 (Loan Start date) and Field89 (Mature date)

Drive a new variable = Field89-Filed88 (values: 13, 16, 20, 30,…)

d) Credit score variables Field27, Field29 and Field30

A=800, A_=750; Character and Nominal ⇒ Numerical, Interval

e) Values "?" and "??" are recorded as blank.

Field 36, 47, 54, 56, 60, 61, 68, 69,71

f) Filed101 (Current loan to value ratio range)

Missing value ⇐ Field 100 (Current loan to value ratio)

Record: "<80"=1, "80-90" =2, …, ">=125"=5

- **Data Partition**

70% train and 30% validation data

- **Data replacement**

Class variables with low levels are replaced by numerical levels.
For example, Field82: “3HH MID MARKET” =3,
Field92 :“MARKET- 'Appraisal May Not Be Required'” =0

- **Data Transformation**

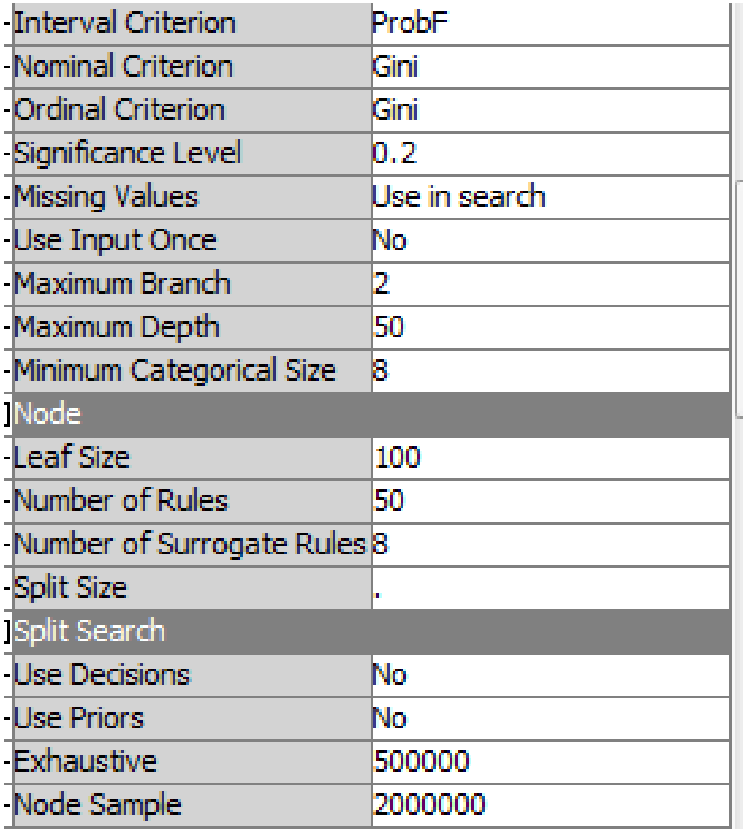
Field93, 109, 116, 117 and so on have high skewness.
Log transformation method

- **Data imputation for missing values**

Tree method for class variables; Mean for interval variables.
Missing value indicators

MODELING APPROACH

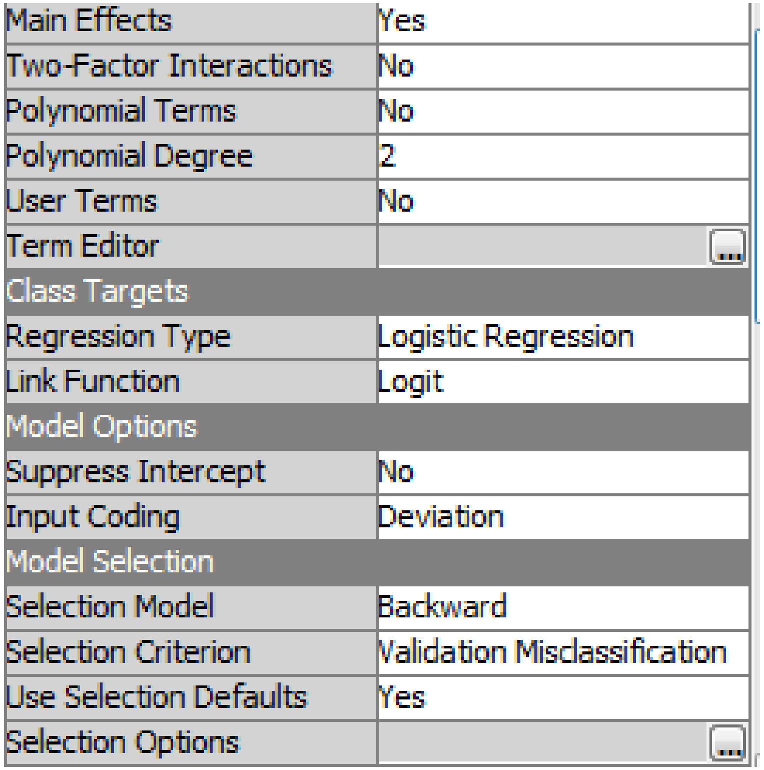
- **Decision tree model**



Interval Criterion	ProbF
Nominal Criterion	Gini
Ordinal Criterion	Gini
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	50
Minimum Categorical Size	8
Node	
Leaf Size	100
Number of Rules	50
Number of Surrogate Rules	8
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	500000
Node Sample	2000000

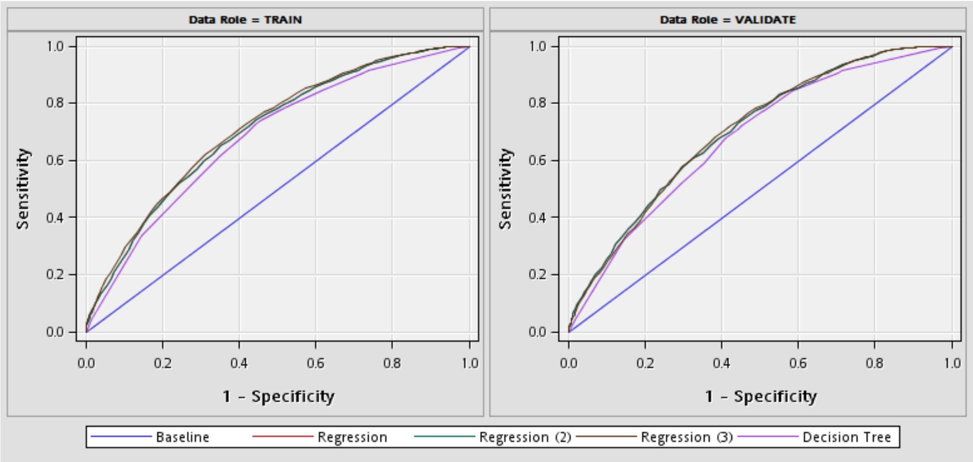
• Logistic Regression model

Stepwise, forward and backward

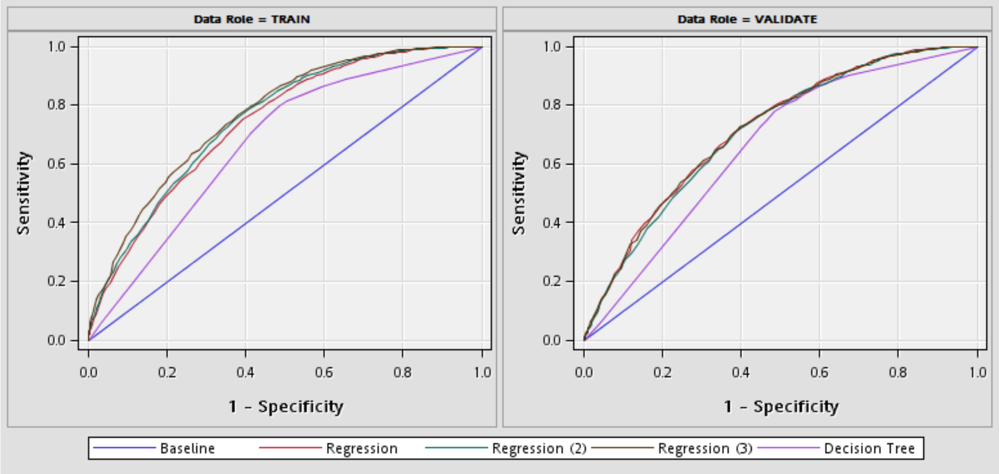


Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes
Selection Options	...

- **Model Comparison**



The Roc Curves for Target1



The Roc Curves for Target2

RESULTS AND CONCLUSIONS

- **Results**

- AUC

	AUC
Target1	0.701
Target2	0.714

- Score the test data

The event probability of each target in sampling is obtained.

Undo sampling:

scoring result = 1/(1+(1/original fraction-1)/(1/oversampled fraction-1)*(1/sampling result-1))

RESULTS AND CONCLUSIONS

- **Conclusions**

- The AUC and probability scores for both targets from EverBank datasets of marketing loan campaign are obtained. The results show the model can accurately forecast customer’s likelihood to respond or get the loan after the next Marketing campaign.

- In the data preparation, oversampling method is used to solve unbalanced dataset problem. Data mining techniques are used to analyze the dataset and prepare the quality data for modeling.

- In the modeling stage, a decision tree model and three logistic regression models are used to model the data. These modeling methods are compared. The best model is a logistic regression model with backward selection.

Thank you!!!

2013 EverBank Cup

Finalist Presentation

STEPHEN JONES

stephen.jones@knights.ucf.edu

SUMMARY

Contest Task

- *Build a predictive model for each target that will accurately forecast customer's likelihood to respond or get the loan after the next Marketing campaign.*

Model Build

- Target1: Backward Regression, ROC .67
- Target2: Backward Regression, ROC .714
- Ideal Candidate: 90% or lower Loan to Value and will save $358 per month in refinancing.

Model Results

- Based on prediction, customers will save $460k per year verses not applying prediction- this is a 59% improvement.
- This customer savings will directly result in higher profitability and growth of EverBank.

Recommendation

DATA ANALYSIS

What is the relationship of the historic variables to Target2?

- CA, GA, FL, WA, AZ have the highest concentration of Target2.
- Target2 has an average Property Value of $208k.

Highest Responses to Target2:

- ‘Sales/Service’ *Occupation*
- ‘Inferred Married’
- ‘No children present’
- *Property value* of ‘$400,000 - $449,999’
- H*ousehold income of* ‘Level 8’ ($100,000 - $124,999)*,* Excludes Level 0(unknown)

Through this analysis, one can begin to get a portrait of a customer that respond for Target2.

MODELING APPROACH

Modeling Target1 & Target2:

- Relatively clean data set and required minimal data preparation.
- Imputed missing variables using the imputation node. I used the mean for missing interval variables. I rejected missing variables with >50%.
- Partitioned data with a 70/30 split.
- I used the following modeling algorithms: Decision Tree, Neural Network, Regression(stepwise), Regression(backward), and Regression(Forward). For Neural Network, I applied MultiLayer Perceptron with 3 hidden units. For regression modeling, I built logistic regression models with logit link function. I also performed an ensemble on the 4 models.

ROC Target1

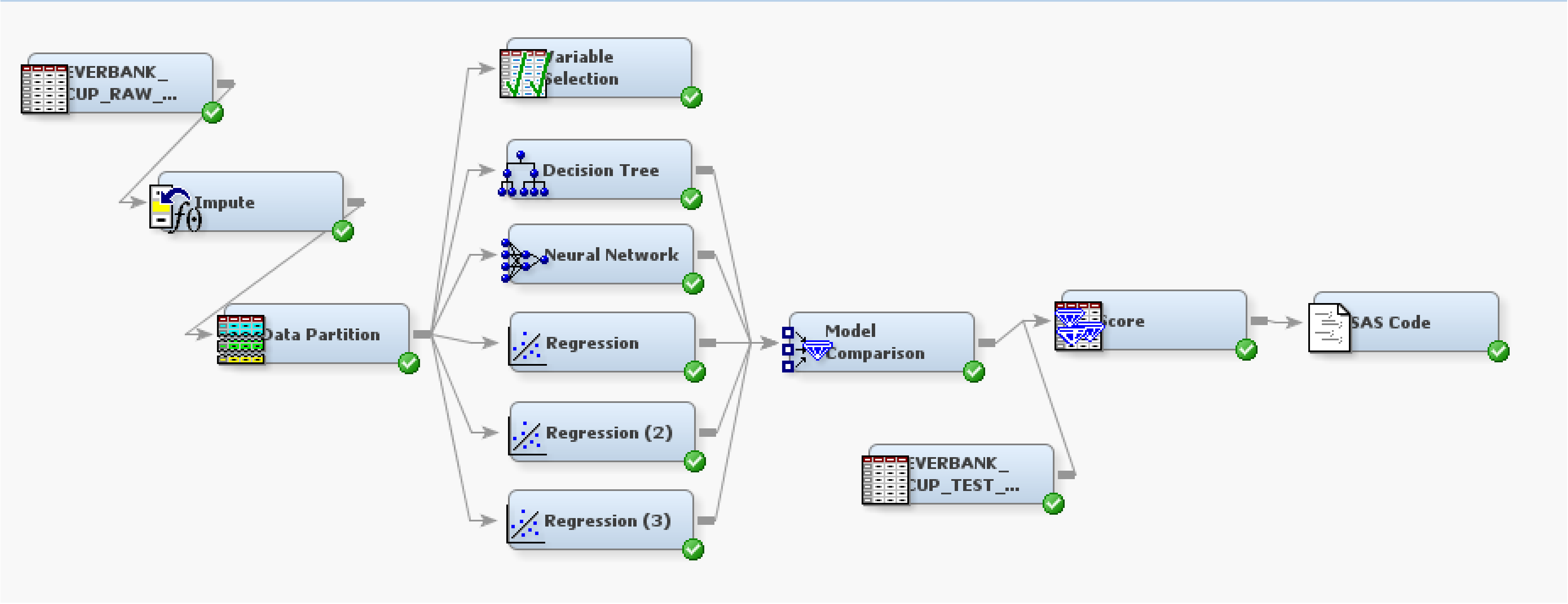
- **Regression(backward): .67**
- Regression(forward): .669
- Regression(stepwise): .668
- Neural Network: .675

ROC Target2

- **Regression(backward): .714**
- Regression(forward): .704
- Regression(stepwise): .707
- Neural Network: .707

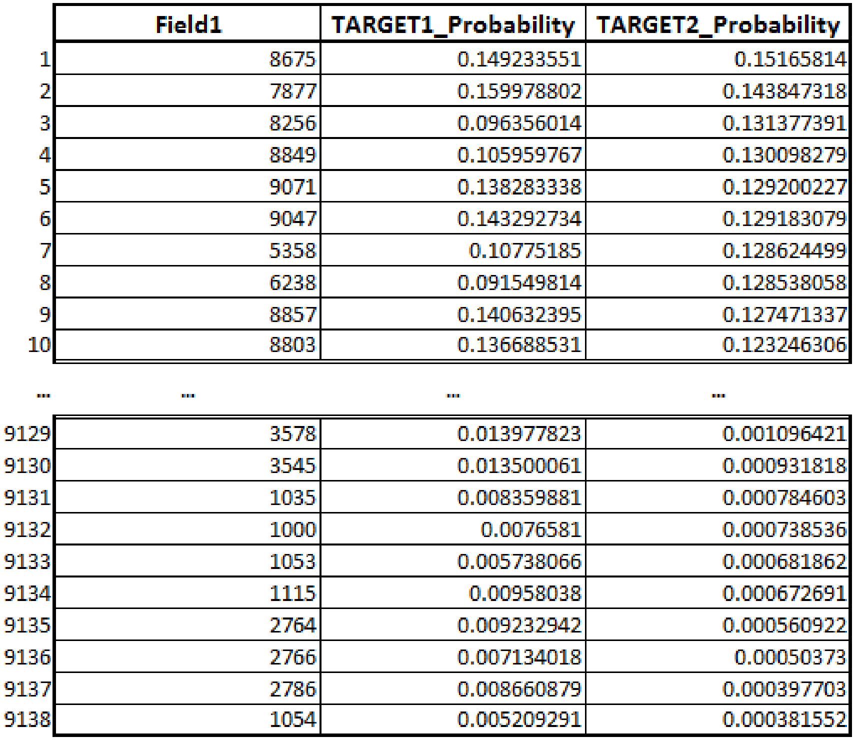
MODELING APPROACH

Sample: *Target1 Enterprise Miner Workflow*



RESULTS

- I built my final model using backward regression.
- I applied a probability factor to each unique field of the test data for Target1 and Target2.



	Field1	TARGET1_Probability	TARGET2_Probability
1	8675	0.149233551	0.15165814
2	7877	0.159978802	0.143847318
3	8256	0.096356014	0.131377391
4	8849	0.105959767	0.130098279
5	9071	0.138283338	0.129200227
6	9047	0.143292734	0.129183079
7	5358	0.10775185	0.128624499
8	6238	0.091549814	0.128538058
9	8857	0.140632395	0.127471337
10	8803	0.136688531	0.123246306
...	...	...	...
9129	3578	0.013977823	0.001096421
9130	3545	0.013500061	0.000931818
9131	1035	0.008359881	0.000784603
9132	1000	0.0076581	0.000738536
9133	1053	0.005738066	0.000681862
9134	1115	0.00958038	0.000672691
9135	2764	0.009232942	0.000560922
9136	2766	0.007134018	0.00050373
9137	2786	0.008660879	0.000397703
9138	1054	0.005209291	0.000381552

Target2 Average:
3.98%

RESULTS: MODEL ASSESSMENT

Financial evaluation:

- *‘Target2 probability factor ‘* X *‘Monthly savings under 30-year loan with new rate’* **=** *‘Expected savings for all customers’*
- Results in $1.2M annual savings

Status quo comparison:

- The average probability of all customers is 3.98% ~*assumed as typical close ratio* ($788k annual savings).
- By using my predictive model, EverBank will achieve 59% ($1.2 M vs. $788k) improvement over a non-predictive approach.
- Valued additional savings of $463k.

RECOMMENDATION

Objective

To optimize effectiveness of marketing plan, company resources, and maximize profit potential.

Recommendation

EverBank pursue the upper 25% Target2 Probability

- Potential customers with greater than 6% probability.
- Of 2253 customers, 83% have positive equity (<1.0 Loan to value).
- **88% improvement** ($744k vs. $394k).

RECOMMENDATION

Optimized Results

- Favorable risk exposure from lender perspective.
- Most cost effective to administer.
- Provides the greatest opportunity for the customers to refinance.

~Leading to improved profitability and growth

QUESTIONS

Thank you

Stephen L. Jones

stephen.jones@knights.ucf.edu

2013 EverBank Cup

Finalist Presentation

RYAN DAGEN

UNIVERSITY OF CENTRAL FLORIDA

STATISTICS DEPARTMENT – DATA MINING PROGRAM

PRESENTED BY: LYNDSEY WEIMER AND AMBER MILLER

DATA EXPLORATION & PREPARATION

- **Initial data set contains 125 variables**
  - 123 predictors and 2 binary targets
  - Target variables represent customer response to previous marketing campaign measured by the filing of loan applications and subsequent acceptance or denial of the loan

- **First step: Studied field definitions to develop better understanding of available data**
  - Looking for variables which stand out as strong predictors of the targets

- **Second step: Utilized graph and stat explore nodes to find relationships with target(s) and possible need for mathematical transformations**
  - Log transformations to increase normality

DATA EXPLORATION & PREPARATION (CONT.)

- **Problems found:**
  - High degrees of “missingness” for several fields
  - Far too many categorical levels for logistic regression processing

- **Addressed these issues using cluster imputation and categorical variable smoothing**
  - Smoothing logit 100 on categorical fields containing >5 levels

- **Lastly: Missing value indicators and missing value pattern variables were created to address potential relationship with targets**

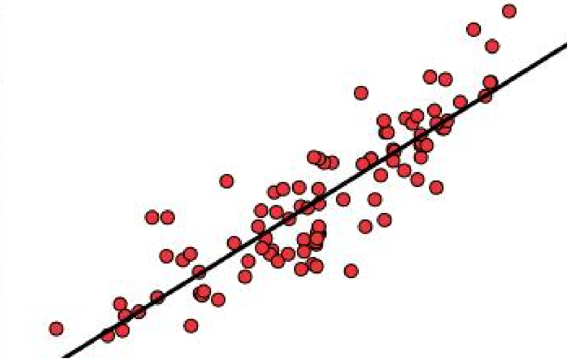
▪ **Note – several modeling iterations were conducted and used to compare preparation techniques. Described above are the final methods.**

DATA MODELING

- **Begin with basic variable selection**
  - Processing limitations
  - Used lax selection criteria

- **Trained model for each target seperatly**

- **Used several different modeling techniques for comparison**
  - Decision Tree (Gradient Boosting)
  - Neural Network
  - Logistic Regression
    - Stepwise selection using interaction and second order terms



- **Selection Criteria**
  - Lowest mean square error from validation set

RESULTS

- **Target1**
  - Logistic Regression
  - Stepwise Selection
  - Too complex with interactive terms, removed with little predictive penalty
  - 12 fields chosen in all, majority from imputation
    - Missing value indicator for Field 37 and logistic transformation for Field 101 also found as significant

- **Highlights**
  - Difference between current interest rate and rate offered by bank
  - New interest rates on both 20 and 30 year loans

- **Performance measures on validation set:**
  - Misclassification Rate – approximately 8.4%
  - Mean Square Error - .075

RESULTS (CONT.)

- **Target2**
  - Again, stepwise Logistic Regression
  - Similarity between targets yields similar selection results
  - 17 fields chosen for “loan acceptance” model
- **Highlights**
  - Ownership of credit card from unknown source by customer
  - Home appraisal necessity indicator
  - More directly speaks to prediction of target2
- **Performance measures on validation set:**
  - Misclassification Rate – approximately 4.5%
  - Mean Square Error - .042

FINAL THOUGHTS

- **Predictive Modeling**
  - With accurate measures, provides ability to score, rank, and prioritize new customer data
  - Constantly expanding client information and customer databases
  - Modeling can optimize future direct marketing efforts using “scientific selection”

- **Modeling on marketing data**
  - Both models can be used to rank future customers displaying similar characteristics of positive past customer behavior
  - Marketing is costly
  - Use model scoring results for Target1 to target marketing areas or individuals with highest propensity to respond to marketing efforts
  - Combine with Target2 to narrow selection to those most likely to complete transaction with institution

Thank you!