Cluster-Based Predictive Modeling on Tax Return Fraud Detection

Tracy Song-Brink

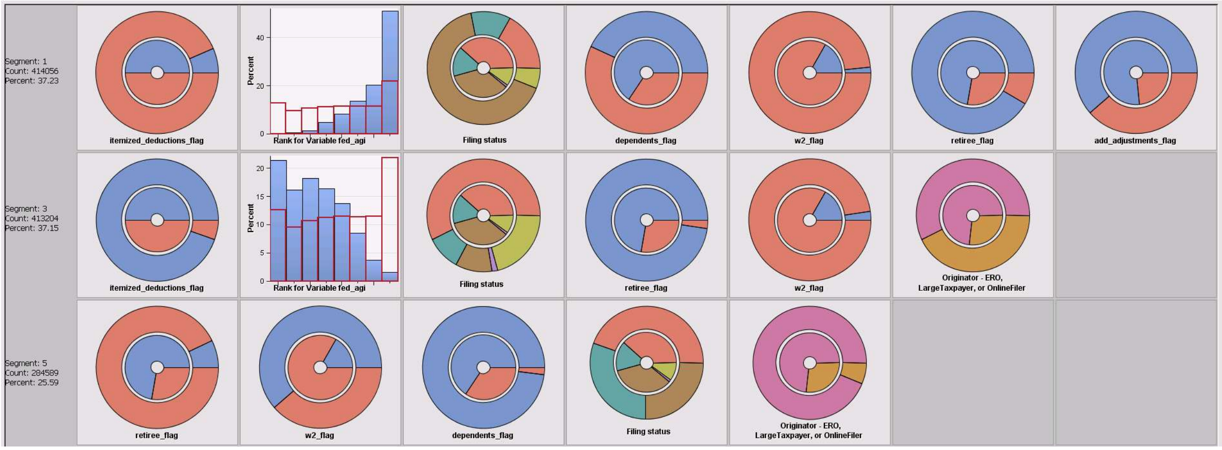
May 15, 2018

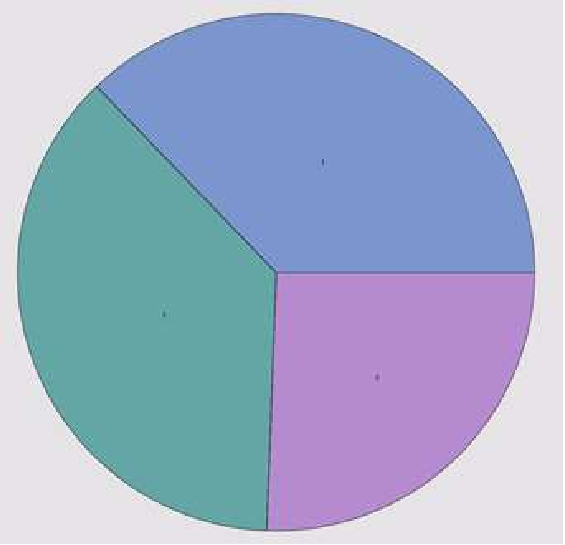
Cluster-Based Predictive Modeling on Tax Return Fraud Detection - Overview

- Goal: build and evaluate segment-based predictive models to detect individual tax return fraud
- Steps:
  - Ran Macros to generate tax payer profile data for analysis
  - Performed clustering on population
  - Built models for each individual segment
  - Selected the best-fitting model for each segment
  - Assessed the performance of segment-based model by comparing the classification rates of the combined segment-based model and population-based model

Cluster-Based Predictive Modeling on Tax Return Fraud Detection - Clustering

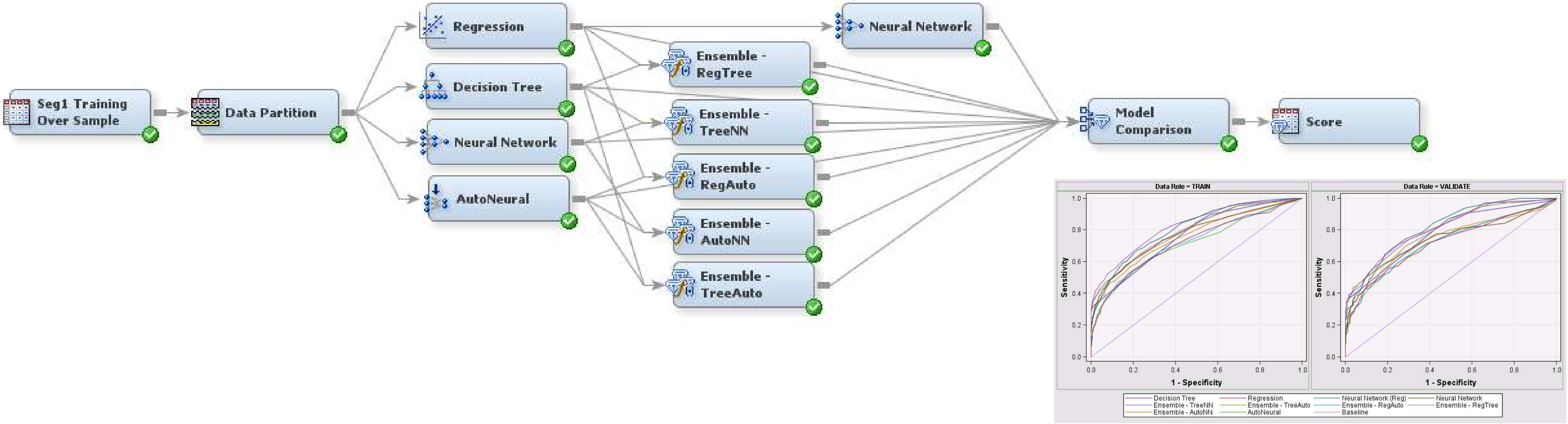
- Ran Macros to generate tax payer profile data for analysis
- Created variables for clustering





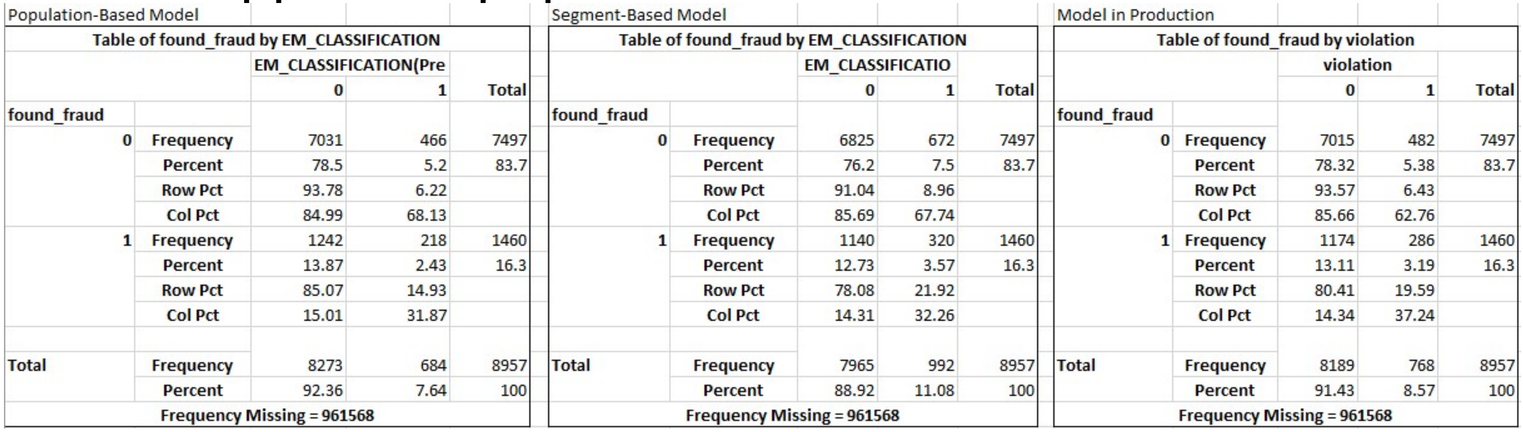
Cluster-Based Predictive Modeling on Tax Return Fraud Detection - Modeling

- Input data split 80/20 for Training/Validation and Test
- Oversampling was performed before modeling
- Selected the model with best ROC and oversampling performance



Cluster-Based Predictive Modeling on Tax Return Fraud Detection - Comparison

- Segment-based model: Models are built based on one individual segment of the data. We applied the score code on test set for the segment. We then merged all segments with predicted value into one dataset for evaluation.
- Population-based model: Models are built based on population and score code is applied to population test set.



Population-Based Model

Table of found_fraud by EM_CLASSIFICATION					
		EM_CLASSIFICATION(Pre			
			0	1	Total
found_fraud					
0	Frequency		7031	466	7497
	Percent		78.5	5.2	83.7
	Row Pct		93.78	6.22	
	Col Pct		84.99	68.13	
1	Frequency		1242	218	1460
	Percent		13.87	2.43	16.3
	Row Pct		85.07	14.93	
	Col Pct		15.01	31.87	
Total	Frequency		8273	684	8957
	Percent		92.36	7.64	100
Frequency Missing = 961568					

Segment-Based Model

Table of found_fraud by EM_CLASSIFICATION				
		EM_CLASSIFICATIO		
		0	1	Total
found_fraud				
0	Frequency	6825	672	7497
	Percent	76.2	7.5	83.7
	Row Pct	91.04	8.96	
	Col Pct	85.69	67.74	
1	Frequency	1140	320	1460
	Percent	12.73	3.57	16.3
	Row Pct	78.08	21.92	
	Col Pct	14.31	32.26	
Total	Frequency	7965	992	8957
	Percent	88.92	11.08	100
Frequency Missing = 961568				

Model in Production

Table of found_fraud by violation				
		violation		
		0	1	Total
found_fraud				
0	Frequency	7015	482	7497
	Percent	78.32	5.38	83.7
	Row Pct	93.57	6.43	
	Col Pct	85.66	62.76	
1	Frequency	1174	286	1460
	Percent	13.11	3.19	16.3
	Row Pct	80.41	19.59	
	Col Pct	14.34	37.24	
Total	Frequency	8189	768	8957
	Percent	91.43	8.57	100
Frequency Missing = 961568				