TAP TO GO BACK TO KIOSK MENU

SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

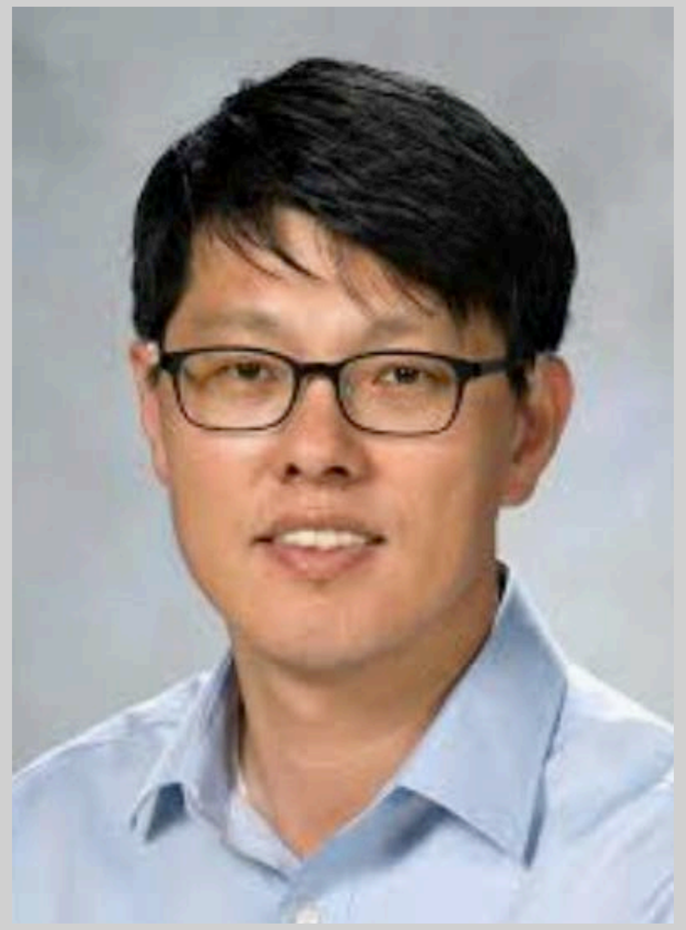
TAP TO GO BACK TO KIOSK MENU

Sunjip Yim M.S., MBA and Ho-Chang Chae, Ph.D.

University of Central Oklahoma

- The study on the predictive modeling for **the adoption of Enterprise Resource Planning (ERP) with business performance** has been lacking so far in large.
- Thus, we answered for this question with massive time-series firm-level data collected by South Korea Statistics agency.
- With more than 11,400 Korean companies' data with 256 variables in each year, we modeled twenty-four **SAS Enterprise (E) Miner** nodes and wrote eight **Python Scikit-learn** programming codes to find the best predictive models for the ERP adoption by firms. During nine years of the survey period, we selected the years of 2006 (the start year), 2010 (the middle year), and 2014 (the end year).
- One of the reasons to conduct this research is to find out the difference in results from SAS E Miner and Scikit-learn.
- We found that there is no fixed best model throughout three separate years. Furthermore, SAS E Miner's best models seem to vary more than in Scikit-learn. At this point. we do not hastily conclude the cause of this phenomenon because, due to the lack of time, our Scikit-learn codes are not exactly identical to the detailed default setting of the well-established SAS E Miner nodes.
- However, even under this best model volatility, the misclassification rates of SAS E Miner and the accuracy of Scikit-learn models surely show the improving tendency as years go by.
- The neural network, (logistic) regression, or random forest method after a precedent variable selection treatment node have a high probability to be the best models for predicting ERP adoption by firms. However, decision trees or support vector machines (SVM) are revealed to be inefficient in predicting ERP adoption.
- In some of the best models, the effect of input variables can be measured. In other best models, we can at least identify which input variables should be treated importantly in other models.

Main Author



Co-Author

Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

Sunjip Yim and Dr. HoChang Chae
University of Central Oklahoma

Intro

- Enterprise Resource Planning (ERP) is one of the most important IT investments, but implementation can be risky.
- What previous research uncovered so far is **what exogenous factors affect ERP adoption.**
- Research needed:
  - **Predictive modeling for the ERP adoption** with various business performances utilizing Machine learning techniques and time-series panel data

Previous Research

- Archive analysis for the ERP adoption (challenges and enablers) including predictive models for the success for the ERP implementation **(Eden et al., 2014)**
- Surveying the ERP adoption with the organization's performance and other factors **(Lorca & de Anders, 2011)**
- Observing firms' positive performance (ROI and ROA) increase only in the third year after the ERP implementation **(Poston & Grabski, 2001)**

Objective

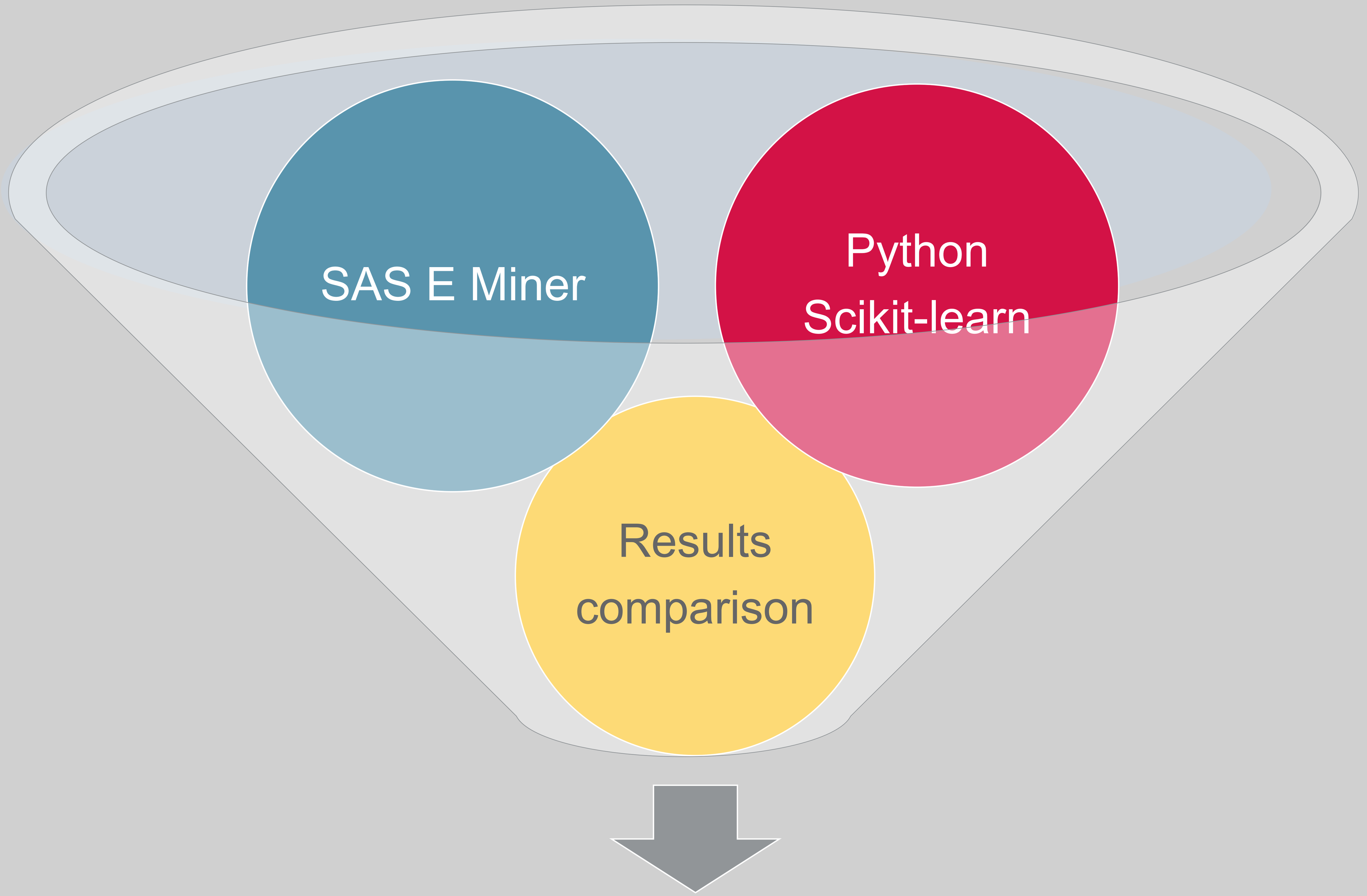
Research Question 1

Among business performance and operating indices, what are the major factors that influence the ERP adoption in the time series data?

Research Question 2

What are the main lessons after conducting and comparing the results from SAS Enterprise (E) Miner and Python Scikit-learn?

9 years' time-series panel data with ERP adoption



Foundlings and lessons

Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

UNIVERSITY OF Central Oklahoma

Sunjip Yim and Dr. HoChang Chae

University of Central Oklahoma

Abstract
Introduction
Methods
Results 1
Results 2
Results 3
Conclusion

Data Source & Description

The Statistics Korea (Gov't of Korea)

URL: http://kostat.go.kr/portal/eng/aboutUs/3/1/index.stati

- Data title: Survey of Business Activities
- Surveyed firms: companies in Korea with at least 50 full-time employees and US$ 0.3 million capital stock
- **Survey period: 9 years (2006 to 2014)**
  - **# of survey variables: 256**
  - **# of rows: 102,743 (11,415 average per year)**

Step 1. Data cleaning with Python

- Treating null values, making dummy variables
- **Total Dataset produced: 12** (= 3 * 2 *2)
  - 3 years (2006, 2010, or 2014) data chosen
  - Standardized or non-standardized (original) data
  - All industry data or manufacturing industry-only data
- We made many sub-datasets; however, for convenience, the result for the 2014 data set including all industry without standardization are mainly dealt as an example.

Step 2. Running models in SAS E Miner and Scikit-learn

- Target variable as **EbizSystem2:**

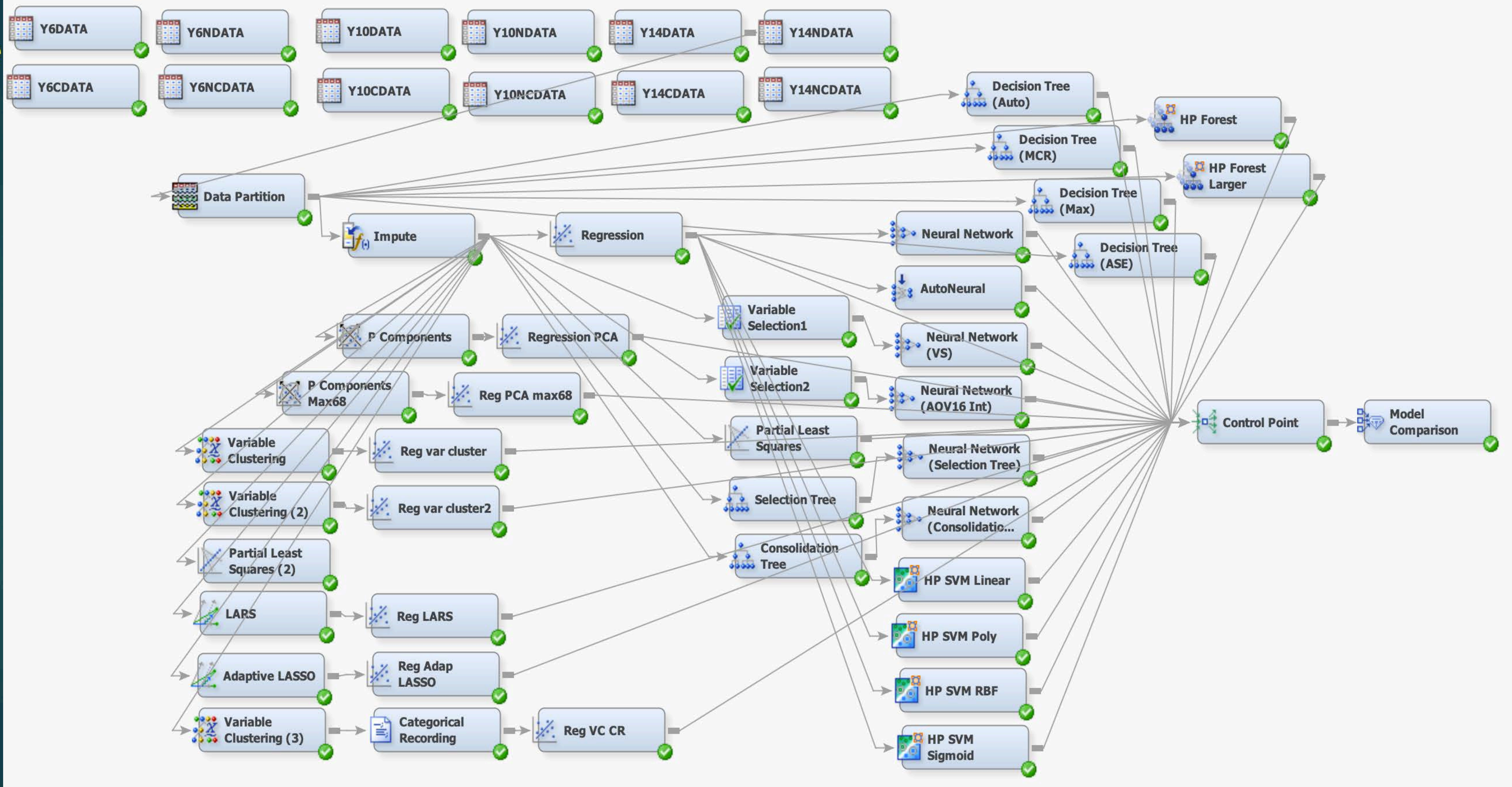
  1 if ERP is adopted or 0 if not.

Variable roles in SAS E Miner

ID	1
Binary input	37
Interval input	77
Nominal input	9
Rejected binary	1
Rejected interval	68
Rejected nominal	10
Rejected unary	53
Target Binary	1

Columns (# of sub-categories): EbizSystem2 (ERP) as target variable

EbizSystem1 to 12 (EbizSsytems in current year)	Liability (3)
EbizSystem2005 1 to 10 (EbizSystems in 2005)	NewEntry (3)
Area	NProfitB4Tax (1)
Assets (9)	Outsourcing (13)
B2B export/import/purchase (7)	Outsourcing cost (1)
Capital (ration) (3)	Overseas (9)
Company ID (ID variable)	ParentCompany (3)
Compensation (4)	RevProft (2)
Cost (17)	R&Dcost (5)
Design (3)	StragegicAlliance (42)
Employment (74)	StockMarketListing (1)
EquityShareCapital (1)	Subsidiary (5)
Franchise (2)	TangibleAsset (8)
IndustryCategory (2)	TradeMark (3)
IntangibleAsset (1)	UtilityModelRight (3)
	Year (1)



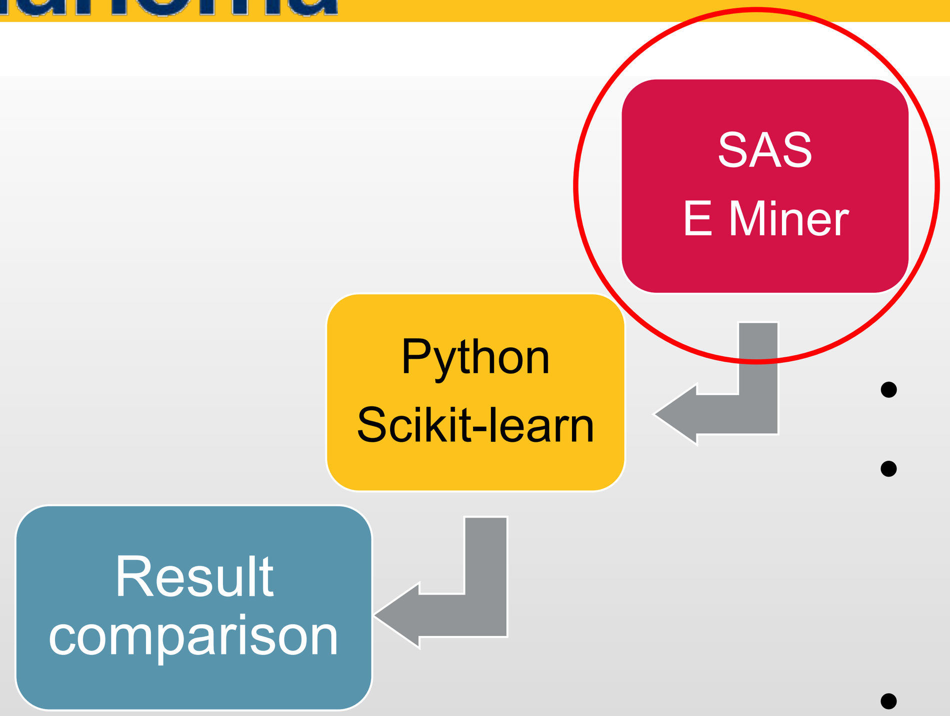
Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

University of Central Oklahoma

Sunjip Yim and Dr. HoChang Chae

University of Central Oklahoma

Abstract
Introduction
Methods
Results 1
Results 2
Results 3
Conclusion



24 models were conducted by SAS E Miner.

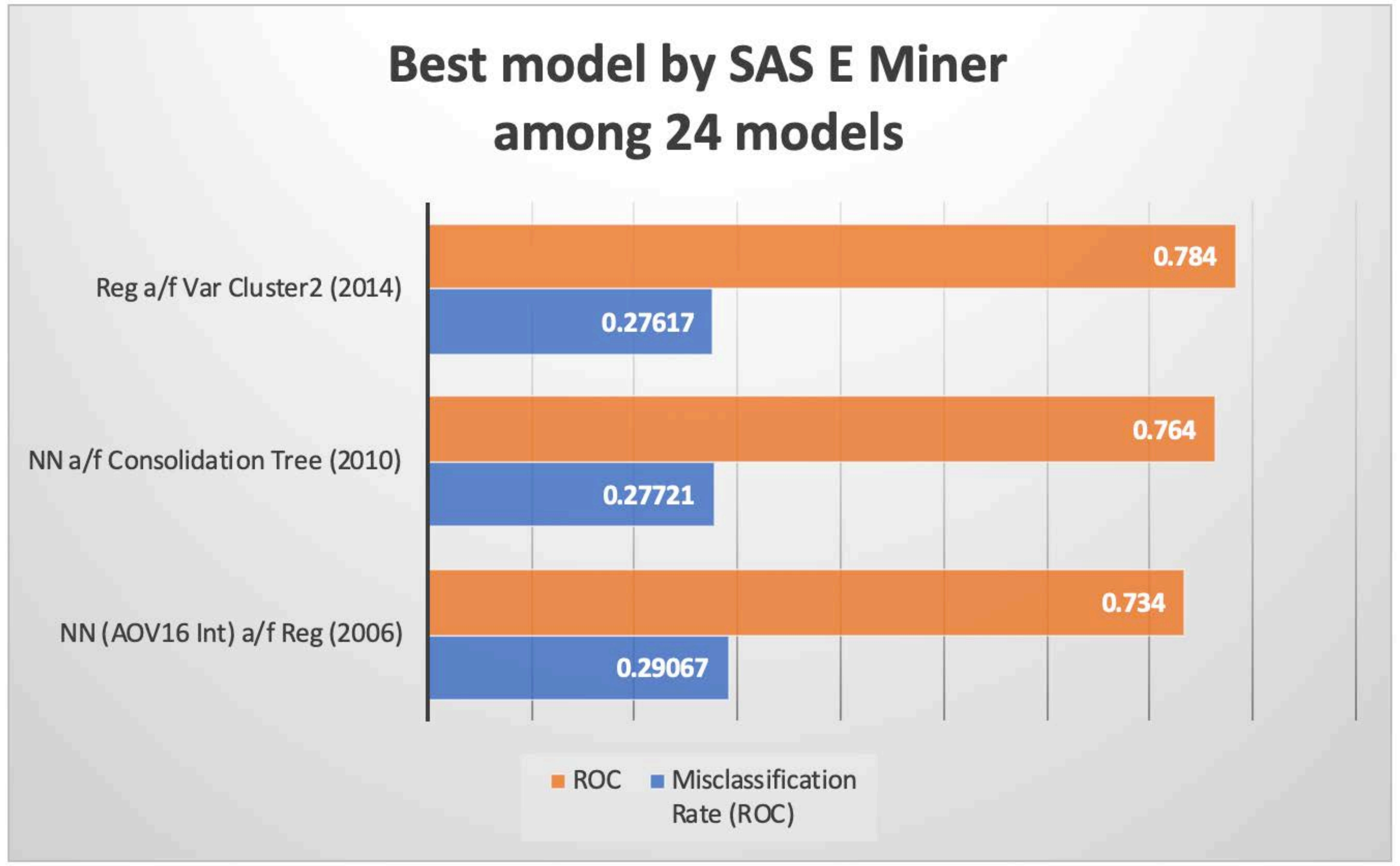
- There is **no fixed best model** in each year to predict the adoption of ERP.
- However, **the neural network, (logistic) regression, or random forest method** after a precedent variable selection treatment node have a high probability to be the best models for predicting ERP adoption.
- Generally, **decision trees or SVM** are proved not to be a good choice in our research.

	Best model's Misclassification Rate	
SAS E Miner <7:3 Partition>	Best Model	Misclassification Rate (ROC)
2006 Data	NN (AOV16 Int) a/f Reg	0.29067 (0.734)
2010 Data	NN a/f Consolidation Tree	0.27721 (0.764)
2014 data	Reg a/f Var Cluster2	0.27617 (0.784)

	2014 Data (12,417 companies)		
SAS E Miner <7:3 Partition>	Rank	Misclassification rate	ROC
NN a/f Consolidation Tree	*	0.27322	0.502
Reg a/f Var Cluster2	1	0.27617	0.784
Neural Network (NN) a/f Reg	2	0.27751	0.785
PLS	3	0.27939	0.783
NN a/f Selection Tree	4	0.27939	0.775
Reg a/f VC CR	5	0.27966	0.786
NN (Auto)	6	0.27966	0.785
Reg	7	0.27993	0.784
Reg a/f Var Cluster	8	0.28046	0.783
HP Forest Larger	9	0.28100	0.792
HP Forest	10	0.28154	0.792
Reg a/f LARS	11	0.28261	0.781
Reg a/f PCA	12	0.28261	0.774
Reg a/f PCA max68	13	0.28261	0.774
NN a/f Var Selection	14	0.28422	0.780
Reg a/f Adap LASSO	15	0.28529	0.779
HP SVM Linear a/f Reg	16	0.28556	0.782
Decision Tree (Auto)	17	0.29334	0.742
Decision Tree (Max)	18	0.29334	0.742
Decsion Tree (MCR)	19	0.29334	0.742
Decision Tree (ASE)	20	0.29898	0.761
NN (AOV16 Int) a/f Reg	21	0.30757	0.754
HP SVM Poly a/f Reg	22	0.32126	0.724
HP SVM Sigmoid a/f Reg	23	0.42834	0.482

* not included in the rank due to too low ROC

	Rank by Misclassification Rate		
SAS E Miner <7:3 Partition>	2006 Data	2010 Data	2014 Data
NN a/f Consolidation Tree	*	1	*
Reg a/f Var Cluster2	18	7	1
Neural Network (NN) a/f Reg	8	2	2
PLS	10	17	3
NN a/f Selection Tree	3	4	4
Reg a/f VC CR	11	18	5
NN (Auto)	0	6	6
Reg	5	10	7
Reg a/f Var Cluster	7	19	8
HP Forest Larger	9	5	9
HP Forest	12	3	10
Reg a/f LARS	2	9	11
Reg a/f PCA	16	20	12
Reg a/f PCA max68	17	21	13
NN a/f Var Selection	4	15	14
Reg a/f Adap LASSO	6	11	15
HP SVM Linear a/f Reg	20	8	16
Decision Tree (Auto)	13	12	17
Decision Tree (Max)	14	13	18
Decsion Tree (MCR)	15	14	19
Decision Tree (ASE)	19	16	20
NN (AOV16 Int) a/f Reg	1	22	21
HP SVM Poly a/f Reg	21	23	22
HP SVM Sigmoid a/f Reg	22	24	23



- Among nine years of observation, we took three sample years: 2006, 2010, and 2014.
- As years pass by and companies adopting ERP increase, the misclassification rate and ROC index are shown to be improved, even though the best model of each year is not fixed.

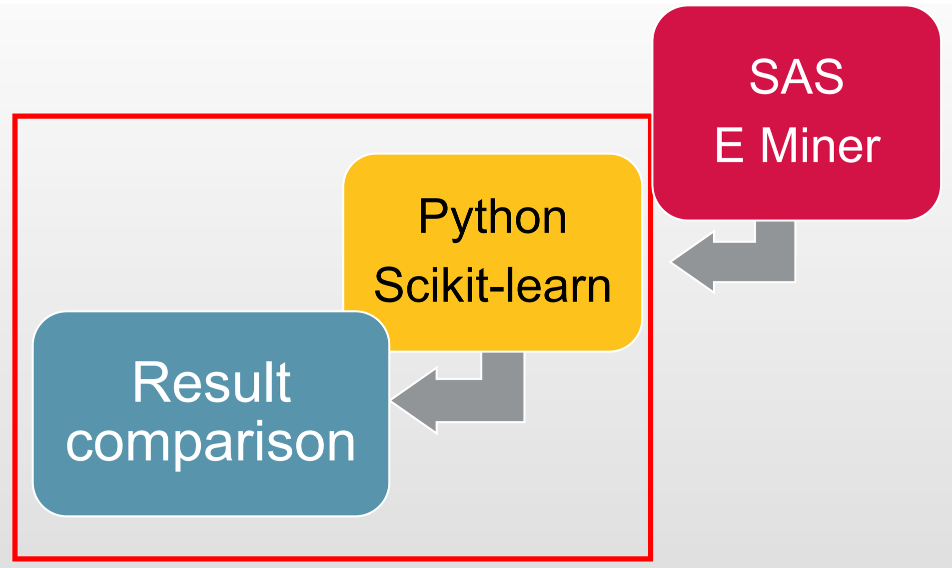
Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

UNIVERSITY OF Central Oklahoma

Sunjip Yim and Dr. HoChang Chae

University of Central Oklahoma

- Abstract
- Introduction
- Methods
- Results 1
- Results 2
- Results 3
- Conclusion



8 models were conducted by both SAS E Miner and Scikit-learn.

- In SAS E Miner, best models in each year seems to vary widely by years.
- Meanwhile, Scikit-learn results a little more stable best models than in SAS E Miner.
- Due to the lack of time, we implemented simpler codes in Scikit-learn than in the SAS E Miner's settings with many options. It may be one of reasons for the stable result in Scikit-learn.

	2014 Data (12,417 companies)					
	SAS E Miner <7:3 Partition>			Scikit-learn <7:3 Partition>		
	SAS Rank	Misclassification rate	ROC	Scikit Rank	(1 - Accuracy)	Accuracy
Reg a/f LARS	5	0.28261	0.781	1	0.27563	0.72437
HP Forest Larger (Random Forest)**	4	0.28100	0.792	2	0.28019	0.71981
Neural Network (NN) a/f Reg	1	0.27751	0.785	3	0.30596	0.69404
Decision Tree	7	0.29334	0.742	3	0.30596	0.69404
NN a/f Selection Tree	2	0.27939	0.775	5	0.34300	0.65700
NN a/f Consolidation Tree	*	0.27322	0.502	6	0.34595	0.65405
HP SVM Linear (or SVM RBF) a/f Reg***	6	0.28556	0.782	7	0.37440	0.62560
Reg	3	0.27993	0.784	8	0.37708	0.62292

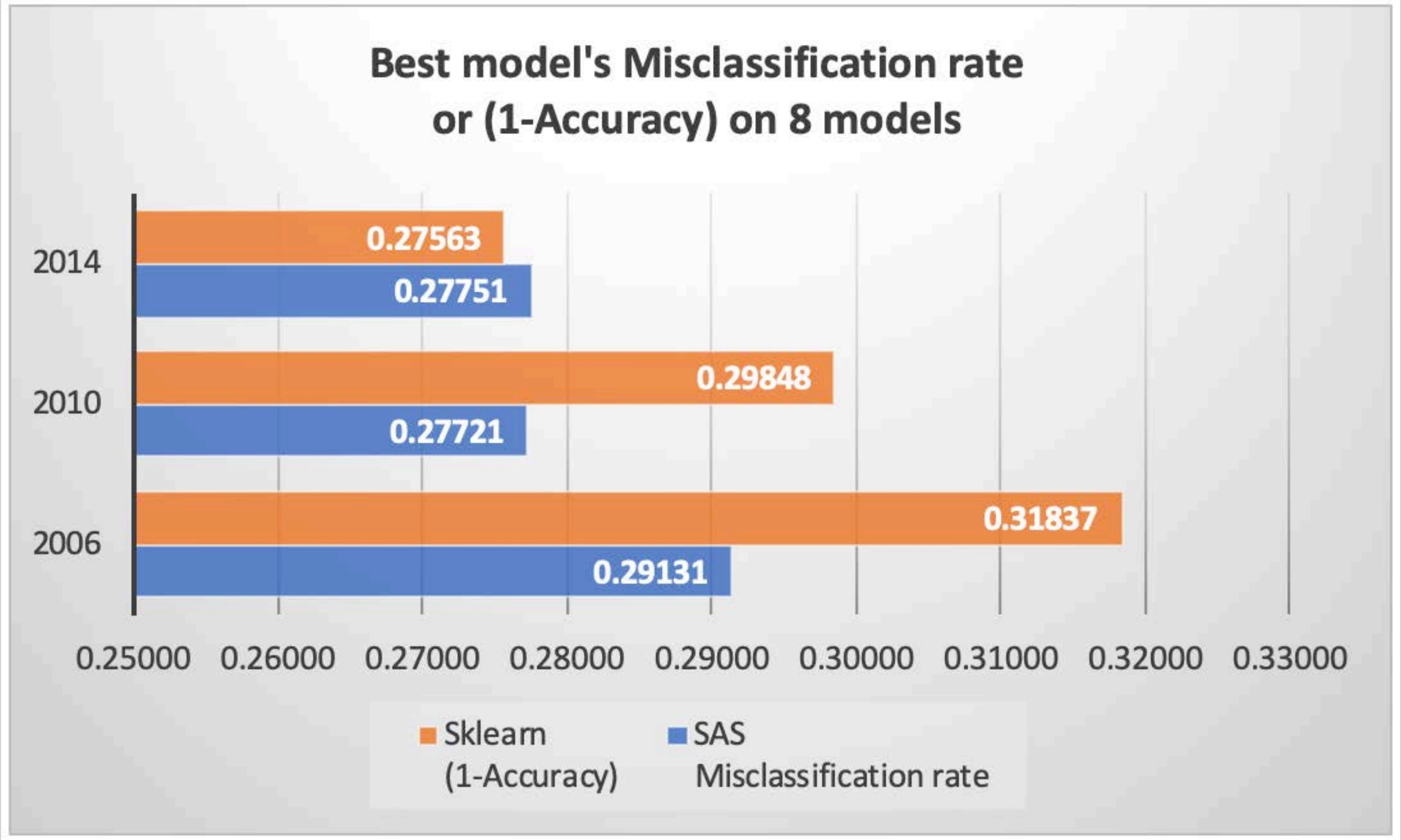
* not included in the rank due to too low ROC

**HP Forest Larger for SAS E Miner and Random Forest for Scikit-learn

***HP SVM Linear for SAS E Miner and SVM RBF for Scikit-learn

	Rankby Misclassification Rate or (1-accuracy)					
<7:3 Partition>	2006 SAS	2006 Sklearn	2010 SAS	2010 Sklearn	2014 SAS	2014 Sklearn
Reg a/f LARS	1	2	6	1	5	1
HP Forest Larger (Random Forest)**	5	1	3	2	4	2
Decision Tree	6	3	8	3	7	3
Neural Network (NN) a/f Reg	4	4	2	4	1	3
NN a/f Selection Tree	2	8	4	7	2	5
NN a/f Consolidation Tree	*	7	1	5	*	6
HP SVM Linear (SVM RBF) a/f Reg***	7	5	5	8	6	7
Reg	3	6	7	6	3	8

	Best Model	Misclassification rate or (1-Accuracy)
2006 SAS	Reg a/f LARS	0.29131
2006 Sklearn	Reg a/f LARS	0.27563
2010 SAS	NN a/f Consolidation Tree	0.27721
2010 Sklearn	Reg a/f LARS	0.29848
2010 SAS	NN a/f Reg	0.27751
2014 Sklearn	Random Forest	0.29848



- Even though best models vary both in SAS E Miner and Scikit-learn, there is a enhancing trend in either misclassification rate or Accuracy.
- Thus, our models can be a starting point to study the factors for ERP adoption in firms afterwards.

UNIVERSITY OF Central Oklahoma

Sunjip Yim and Dr. HoChang Chae

University of Central Oklahoma

The effect of input variables on the adoption of EPR (target variable)

- For the model of regression after LARS (Best model) in 2014 data, below are selected input variables in the (logistic) regression: Compensation3, Compensation4, EBizSystem10, EBizSystem3, EBizSystem5, EBizSystem6, IndCategory2, M_Asset3, M_Asset9, M_B2B_purchase1, Outsourcing1, Outsourcing10, Outsourcing11, Outsourcing2, Outsourcing3, Outsourcing7, Outsourcing8, ParentCompany1, StockMktListing, and Subsidiary1 where M_Variable means the imputation indicator.

- The interpretation of the effect of the input variables on the target variable can be checked on the odd ratio table provided in the SAS E Miner result.

- Part of the odds ratio table:

Odds Ratio Estimates

Effect		Point Estimate
Compensation3	1 vs 5	0.704
Compensation3	2 vs 5	1.216
Compensation3	3 vs 5	1.029
Compensation3	4 vs 5	0.988
Compensation4	1 vs 5	0.624
Compensation4	2 vs 5	1.064
Compensation4	3 vs 5	0.688
Compensation4	4 vs 5	0.829
EBizSystem10	0 vs 1	5.383
EBizSystem3	0 vs 1	0.440
EBizSystem5	0 vs 1	0.486
EBizSystem6	0 vs 1	1.414
IndCategory2	1 vs 96	0.290

- For the model of neural network after regression (one of the top models) in 2014 data, below are selected input variables in the (logistic) regression before the Neural Network node: Compensation3, Compensation4, EBizSystem10, EBizSystem3, EBizSystem5, EBizSystem6, EBizSystem8, IMP_OutsourcingCost, IMP_TAssetC3, IMP_emp3, IndCategory2, M_Asset3, M_Asset9, M_B2B_purchase1, M_RNDcost1, Outsourcing1, Outsourcing10, Outsourcing11, Outsourcing2, Outsourcing3, Outsourcing7, Outsourcing8, ParentCompany1, StockMktListing, and Subsidiary1 where IMP_Variable and M_Variable mean the imputated variable and the imputation indicator each.

- The above input variables are fed into the neural network node right after the (logistic) regression node.

- As you know well, it is hard to interpret the weights of input variables on the neural network model.

- However, at the practical level, we can confirm which input variables on the whole data should be selected and fed into the neural network model here.

Lessons for the policy-practitioners

- We can interpret the effect of input variables on the adoption of ERP on some best models or cannot on others due to the characteristics of neural network models.
- However, at least, there may be a great possibility for us to find which factors should be on the best models. Therefore, those variables should be carefully treated by policy makers.

Machine Learning Data Analysis for the ERP Adoption and Enterprise Performance with SAS® Enterprise Miner and Python Scikit-learn

Sunjip Yim and Dr. HoChang Chae

University of Central Oklahoma

Foundlings

- The best model of each year vary while data standardization does not impact the overall analysis result.
- The neural network, (logistic) regression, or random forest method after a precedent variable selection treatment node have a high probability to be the best models for predicting ERP adoption. However, decision trees or SVM turns out to be inefficient for this role.
- SAS E Miner's best models vary more than in Scikit-learn.
- In some best models, the effect of input variables can be measured. Otherwise, we can at least identify which input variables should be treated importantly in other models.
- Throughout the nine years of the observation period, the misclassification of SAS E Miner models and the accuracy of Scikit-learn models have an improving trend as years go by.

Further research

- The gradient boosting method in Scikit-learn shows to yield the best model all through the three years selected (i.e., 2006, 2010, and 2014). If it is truly so for the remaining six years, then the reason for that should be worth for being searched for.
- Only less than half of the models in SAS E Miner can be easily coded in Scikit-learn. The remaining models demand too much time and effort in developing the codes. Especially, the programming codes in Scikit-learn for inputting variables resulted from the precedent variable selection process into a new node should be carefully developed if needed.

References

Eden, R., Sedera, D., & Tan, F. (2014). Sustaining the momentum: archival analysis of enterprise resource planning systems (2006–2012). *Communications of the Association for Information Systems*, *35*(1), 3.

Grabski, S. V., Leech, S. A., & Schmidt, P. J. (2011). A review of ERP research: A future agenda for accounting information systems. *Journal of information systems*, *25*(1), 37-78.

Lorca, P., and J. de Andrés (2011) "Performance and Management Independence in the ERP Implementations in Spain: A Dynamic View", *Information Systems Management,* (28)2, pp. 147–164.

Poston, R., and S. Grabski. 2001. Financial impacts of enterprise resource planning implementations. *International Journal of Accounting Information Systems* 2: 271–294.

Ram, J., Corkindale, D., & Wu, M. L. (2013). Enterprise resource planning adoption: structural equation modeling analysis of antecdants. *Journal of Computer information systems*, *54*(1), 53-65.

Ranganathan, C., and C.V. Brown (2006) "ERP Investments and the Market Value of Firms: Toward an Understanding of Influential ERP Project Variables", *Information Systems Research,* (17)2, pp. 145–161.

MARCH 29 – APRIL 1 | WASHINGTON, DC | #SASGF