• 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.
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

Abstract

Methods

Results 1

Results 2

Results 3

Conclusion

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)

Objectives

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?
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.
Abstract

Introduction

Methods

Results 1

Results 2

Results 3

Conclusion

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

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.

Table 1: Summary of the best models in each year

<table>
<thead>
<tr>
<th>Year</th>
<th>Best Model</th>
<th>Misclassification Rate</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006 Data</td>
<td>NN (AV16 Int) a/f Reg</td>
<td>0.29067 (0.734)</td>
<td></td>
</tr>
<tr>
<td>2010 Data</td>
<td>NN a/F Consolidation Tree</td>
<td>0.27721 (0.764)</td>
<td></td>
</tr>
<tr>
<td>2014 data</td>
<td>Reg a/F Var Cluster2</td>
<td>0.27617 (0.784)</td>
<td></td>
</tr>
</tbody>
</table>

Results 3

- 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.
Abstract

Introduction

Methods

Results 1

Results 2

Results 3

Conclusion

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

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg a/f LARS</td>
<td>0.28261</td>
<td>0.27981</td>
<td>0.28019</td>
<td>0.30596</td>
<td>0.37708</td>
<td>0.62292</td>
</tr>
<tr>
<td>HP Forest Larger (Random Forest)**</td>
<td>0.28100</td>
<td>0.27751</td>
<td>0.27960</td>
<td>0.30596</td>
<td>0.69404</td>
<td>0.69404</td>
</tr>
<tr>
<td>Neural Network (NN) a/f Reg</td>
<td>0.29334</td>
<td>0.7432</td>
<td>0.30596</td>
<td>0.69404</td>
<td>0.69404</td>
<td>0.69404</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.27939</td>
<td>0.7755</td>
<td>0.34300</td>
<td>0.65700</td>
<td>0.37708</td>
<td>0.62292</td>
</tr>
<tr>
<td>NN a/f Consolidation Tree</td>
<td>0.27322</td>
<td>0.5026</td>
<td>0.34595</td>
<td>0.65405</td>
<td>0.62560</td>
<td>0.62560</td>
</tr>
<tr>
<td>HP SVM Linear (or SVM RBF) a/f Reg***</td>
<td>0.28556</td>
<td>0.7802</td>
<td>0.37440</td>
<td>0.62560</td>
<td>0.62560</td>
<td>0.62560</td>
</tr>
<tr>
<td>Reg</td>
<td>0.27939</td>
<td>0.7848</td>
<td>0.37708</td>
<td>0.62292</td>
<td>0.62292</td>
<td>0.62292</td>
</tr>
</tbody>
</table>

* 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

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.
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, EBizSystem3, EBizSystem5, 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:

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compensation3 1 vs 5</td>
<td>0.704</td>
</tr>
<tr>
<td>Compensation3 2 vs 5</td>
<td>1.216</td>
</tr>
<tr>
<td>Compensation3 3 vs 5</td>
<td>1.029</td>
</tr>
<tr>
<td>Compensation3 4 vs 5</td>
<td>0.988</td>
</tr>
<tr>
<td>Compensation4 1 vs 5</td>
<td>0.624</td>
</tr>
<tr>
<td>Compensation4 2 vs 5</td>
<td>1.064</td>
</tr>
<tr>
<td>Compensation4 3 vs 5</td>
<td>0.688</td>
</tr>
<tr>
<td>Compensation4 4 vs 5</td>
<td>0.829</td>
</tr>
<tr>
<td>EBizSystem10 0 vs 1</td>
<td>5.383</td>
</tr>
<tr>
<td>EBizSystem3 0 vs 1</td>
<td>0.440</td>
</tr>
<tr>
<td>EBizSystem5 0 vs 1</td>
<td>0.486</td>
</tr>
<tr>
<td>EBizSystem6 0 vs 1</td>
<td>1.414</td>
</tr>
<tr>
<td>IndCategory2 1 vs 96</td>
<td>0.290</td>
</tr>
</tbody>
</table>

- 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 imputed 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.
Abstract

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.

References


