Applying Ensemble Machine Learning Algorithms to Predict Professional Career Development Preferences among University Students

Galo Paiva and Carlos Hernández
Departamento de Ingeniería Industrial y Sistemas
Universidad de La Frontera
Temuco, Chile
galo.paiva@ufrontera.cl, carlosalberto.hernandez@ufrontera.cl

Abstract

This research is focused on the development and comparison of models based on ensemble machine learning algorithms to predict professional career development preferences among students five years after their graduation using the results of the University Entrepreneurial Spirit Students' Survey 2018. The research is carried out following a classic 4-stage methodology (analysis, design, development, and validation). During the analysis, surveys are thoroughly reviewed and preprocessed. During the design, questions are grouped and combined to build 11 predictive models. Construction and validation are carried out entirely using the software WEKA. For the purposes of this investigation 1.121 surveys are considered. Initially the dataset is split up in a subset for training and test (80%) and a subset for validation (20%). The approach to predict students' mid-term career preferences comprised the use of an ensemble scheme (stacking) composed by a logistic regression as meta-model, and a decision tree, and a support vector machine as base models. Experimental results show that half the proposed models predict correctly around 77% of the surveys. In conclusion, ensemble models can be used to predict students' professional career development preferences. However, predictions' accuracy depends on the attribute selection.

Keywords

Machine Learning, Ensemble Algorithms, Predictive Model, Cross-Validation, GUESSS 2018

1. Introduction

After college, newly graduated professionals face many challenges in the job market. An important decision to be made is whether to apply for a job in a company or to start their own business. The factors involved in this decision have been discussed by several authors. Some based their argumentation on the theory of planned behavior (Ajzen, 1991), others on the theory of the human motivation (McClelland, 1987).

There are multiple factors that can explain the entrepreneurial attitude. Usually, they are classified in two categories. On the one hand are the personal characteristics such as risk aversion, autonomy, goal-orientation, power, and management knowledge. On the other hand, factors related to social environment such as family's entrepreneurial background, economy, formal education and training, and social networks (Sieger, 2011; Zellweger et al., 2011; Bagheri et al., 2015; Gorgievski et al., 2018; Israr and Saleem, 2018; Bosma et al., 2021).

During their university years, students acquired knowledge and develop specific skills that are required to perform properly during their future professional. These skills can be classified as cognitive, socio-emotional, and technical (Prada and Rucci, 2016). Students can also develop entrepreneurial ecosystems whose success depends on the connectedness and effective filtration, and on having a strong local and interregional character (Prokop, 2021). Besides that, the orientation of the university education plays an important role too (Franke and Lüthje, 2004).

The Global University Entrepreneurial Spirit Students' Survey (GUESSS) is a global study conducted every 2-3 years at universities in more than 50 countries by University of Saint Gallen (Switzerland). Its main objective is to determine the factors that drive students' entrepreneurial intentions and activities by means applying a model proposed by Sieger, et al. (2011), which states that the entrepreneurial intention depends on the behavioral attitude, subjective norms, and

perceived behavioral control. Being the entrepreneurial personal background, the motivations, and the entrepreneurial family background the influential variables.

1.1 Objective

To apply machine learning techniques to predict professional career development preferences among university students five years after graduation by means of building predictive models based on ensemble algorithms.

2. Literature Review

2.1 Machine learning

Machine learning can be seen as a branch of artificial intelligence (AI). It comprises several methodologies, techniques, and algorithms to find patterns and to learn from datasets through experience. There exist supervised, unsupervised, and reinforcement learning algorithms. In supervised learning, the training is carried out using datasets that contain the class to be predicted. In unsupervised learning, instead, the desired class is not known.

2.2 Classification, prediction, and forecasting

There are several important concepts in machine learning. Classification can be understood as the determination of the class with a nominal value in an unseen dataset using a previously trained model. Prediction is the estimation of a numeric value for the desired dependent variable. Forecasting, on the other hand, is the prediction of future values using time series.

2.3 Hold out and cross-validation

In machine learning, holding out refers to the split up of a dataset into a set for training and a set for testing. The test dataset help assess the performance of the predictive or classification model on unseen data. Commonly the split proportion is 80% for training and 20% for testing. Cross-validation, on other hand, refers to the random split up of a dataset into k folds. During the building, k-1 folds are used for training while the left one is used to test model's performance. Training and testing are repeated iteratively k times until all folds have been used for training and for testing (Figure 1). The goal is to minimize the risk of the overfitting that could occur when holding out. In the case of a cross-validation, results are different in each iteration because the folds selected for training and testing have been interchanged. The k results are finally averaged.

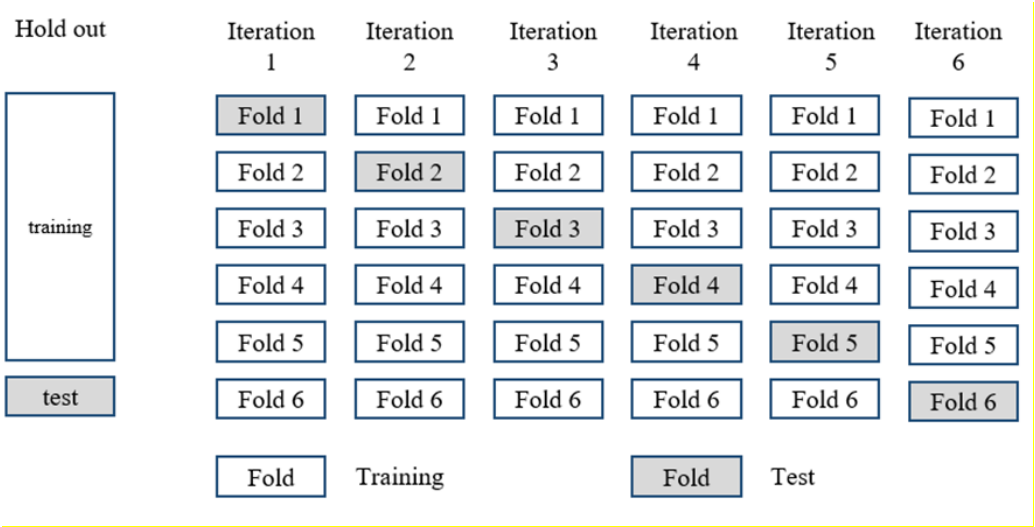

Figure 1. Hold out and cross-validation (k=6)

2.4 Overfitting and generalization

Overfitting occurs when a model learns from the training dataset so well that it does not have a good performance when tested on an unseen dataset. The model does not generalize due to the incorporation of details from the training data that will be easily found in new data (Figure 2).

Generalization is the ability of a trained model to have a good performance on unseen data, which is the objective of building classification models. The usefulness of the model depends on that.

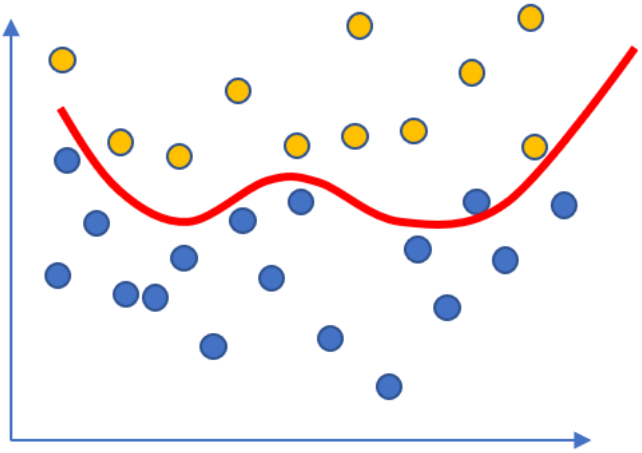

Figure 2. Overfitting

2.5 Replication

In statistics, replication is repetition of an experiment under similar conditions to estimate the variability of phenomenon under study. When using cross-validation, the partitioning of the dataset in k folds depends on a specific seed number. Since different seeds produce different folds, the results are different after each iteration. The mean and the standard deviation can be estimated and analyzed afterwards.

2.6 Meta-learning and ensemble algorithms

Meta-learning, or learning to learn, is the use of a learning algorithm to learn from the predictions made by other learning algorithms. Basically, it combines the predictions of several machine learning algorithms to make new predictions. The so-called ensemble machine learning algorithms present a multi-level structure to carry out learning tasks. There is a meta-algorithm (level 1) that leans from the predictions made by several base algorithms (level 0). The ensemble usually can predict better than any of its single algorithms. There are several ensemble algorithms, being the stacking one of the most widely used (Figure 3).

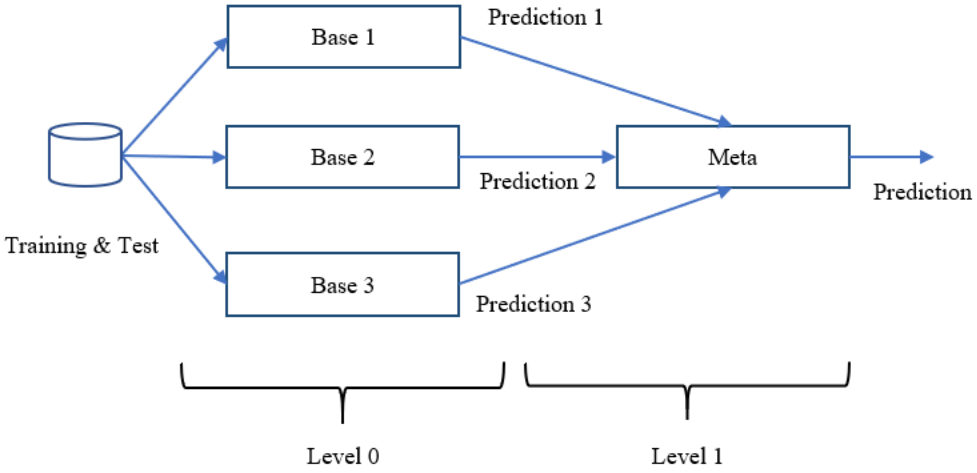

Figure 3. Ensemble algorithm

3. Methods

The present research is carried out following a classic 4-stage model: analysis, design, construction, and validation (Figure 4).

Figure 4. Four-stage model

3.1 Analysis

During the analysis, a complete review of the answers to the GUESSS 2018 surveys conducted at Universidad de La Frontera (Temuco, Chile) is carried out. A total of 1.121 surveys are considered, all of which are properly pre-processed for the subsequent stages. GUESSS 2018 questionnaire is organized in several domains. Most of the answers are expressed in simplified scales, from 1 to 5 or from 1 to 7. Some of these questions are selected to be the attributes of the classification models proposed in this work, leaving the question related to the development career preference five years after graduation as the attribute to be predicted. A total of 57 questions grouped in 7 domains are selected for this research (Table 1).

Table 1. Selected domains from GUESSS 2018

	Domain	**Questions**
D.1	Subject	7
D.2	Career preferences	2
D.3	University environment	13
D.4	Entrepreneurial interests	24
D.5	Entrepreneurial family background	2
D.6	Social environment	6
D.7	Personal information	3

3.2 Design

The dataset, a matrix of 1.121 rows (instances) by 58 columns (57 attributes and the class), is split up to create 2 subsets. The first one for training and test, contains 80% of the data (900 instances). The remaining 20% of the data is left in a separate dataset to be used during the validation stage (221 instances). An instance can be understood as a row containing all the answers of a single survey.

To identity the relevance of each domain on the classification result, different combinations of questions are defined to build and compare 11 classification models (Table 2).

Table 2. Classification models' design

		M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
D.1	Subject	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
D.2	Career preferences	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
D.3	University environment	✓		✓	✓	✓	✓		✓		✓	✓	✓
D.4	Entrepreneurial interest	✓	✓		✓	✓			✓	✓		✓	✓
D.5	Business family background	✓	✓	✓		✓		✓	✓	✓	✓		✓
D.6	Social environment	✓	✓	✓	✓			✓	✓	✓	✓	✓	
D.7	Personal information	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The resulting classification models are compared by means of the percentage of correct predictions using data in the validation dataset. Additionally, curves Precision-Recall and the area under the curve ROC (ROC AUC) are considered for comparison too (Davis and Goadrich, 2006).

When dealing with classification problems it is important to consider the class balance. In the case of heavily imbalanced dataset, the performance metrics should include the Precision-Recall curves along with ROC AUC (Saito and Rehmsmeier, 2016). For the purposes of this investigation, based on preliminary trial and error experiments with the training and test dataset, three learning algorithms are selected: logistic regression, decision tree, and support vector machine. The ensemble scheme selected is the well-known stacking (Table 3)

Table 3. Ensemble algorithm's configuration

Ensemble	Meta-algorithm (level 1)	Base algorithms (level 0)
Stacking	Logistic regression	Decision tree
		Support vector machine

3.3 Construction

As aforementioned, the goal is to predict whether the students' career development preference five years after graduation is to be running their own business or to be working in a company. The desired class can take the two values: business owner or employee. There is also a third value for those ones who do not have any preference yet. All the models presented in this work are built using the data mining software WEKA (Witten et al., 2017). Initially, all models are trained and tested applying a cross-validation scheme of k=10 folds on a dataset of 900 instances (Table 4).

Table 4. Models' performance on training and testing dataset with cross-validation k=10 (900 instances)

		Weighted Average		
Model	%	ROC AUC	Precision	Recall
M0	77.22	0.847	?	0.772
M1	77.22	0.855	?	0.772
M2	51.89	0.585	?	0.519
M3	77.22	0.846	?	0.772
M4	77.22	0.847	?	0.772
M5	49.11	0.538	?	0.491
M6	51.00	0.562	?	0.510
M7	47.67	0.539	?	0.477
M8	48.89	0.554	0.543	0.489
M9	77.11	0.847	?	0.771
M10	50.11	0.553	?	0.501
M11	52.56	0.576	?	0.526

Table 5. Models' performance on training and testing dataset with k=10 cross-validation and 10 replications

Model	Average Correct Predictions (%)	Standard Deviation (10 replications)
M0	77.18	3.07
M1	77.23	3.03
M2	51.16	4.35
M3	77.17	2.99
M4	77.07	3.14
M5	48.69	5.00
M6	50.87	5.08
M7	50.39	4.62
M8	48.18	4.58
M9	76.92	2.93
M10	48.18	5.14
M11	51.38	5.00

In some cases, it is not possible to calculate the values in the column Precision because some of the classes has no instances. This is shown later in the confusion matrices (Table 9). Although cross-validation helps reduce the risk of overfitting, the effect of the fold partitioning remains (Powers and Atyabi, 2012). A simple solution is the replication of the experiments with a different fold partitioning each time. For the purposes of this research 10 replications are run, which means that each model is trained and tested 100 times. The results show that the standard deviation of the replications fluctuates between 3% and 5% (Table 5).

3.4 Validation

The validation of the trained and tested ensemble learning models is carried out with the validation dataset held out during the stage of analysis. This dataset contains unseen 221 instances or rows, equivalent to 20% of the data.

4. Data Collection

The validation results show that approximately half of the models can generalize relatively well on unseen data (Table 6). In all cases the percentage of correct classification is consistent with the percentage obtained with the training dataset (Table 5).

Table 6. Models' performance with the validation dataset (221 instances)

Model	Attributes	Correct Predictions (%)	ROC AUC	Precision	Recall
M0	57	77.83	0.875	?	0.778
M1	44	77.83	0.862	?	0.778
M2	33	44.34	0.532	?	0.443
M3	55	77.83	0.879	?	0.778
M4	51	77.83	0.864	?	0.778
M5	25	44.34	0.573	?	0.443
M6	13	51.13	0.575	?	0.511
M7	12	47.51	0.568	?	0.475
M8	18	43.89	0.526	?	0.439
M9	29	77.83	0.864	?	0.778
M10	7	46.15	0.537	?	0.462
M11	11	49.32	0.569	?	0.493

5. Results and Discussion

The proposed models have significant differences in performance, which indicates that the selection of questions (attributes) is relevant for the investigation (Table 7).

Table 7. Models' size (attributes)

Model	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
Attributes	57	44	33	55	51	25	13	12	18	29	7	11

5.1 Numerical Results

Confusion matrices are useful to summarize the prediction results in tables. The diagonal of the matrix contains the number of instances correctly classified. The other cells present incorrect classifications (Table 8).

Table 8. Confusion matrix

Class 1	Class 2	Class 3
Instance class 1 classified **correctly** as class 1	Instance class 1 classified **incorrectly** as class 2	Instance class 1 classified incorrectly as class 3
Instance class 2 classified **incorrectly** as class 1	Instance class 2 classified **correctly** as class 2	Instance class 2 classified **incorrectly** as class 3
Instance class 3 classified **incorrectly** as class 1	Instance class 3 classified **incorrectly** as class 2	Instance class 3 classified **correctly** as class 3

The results for both datasets, training-test (900 instances) and validation (221 instances), show that the models M0, M1, M2, M3, M4 and M9 classify around 77% of the instances correctly (Table 9).

Table 9. Resulting confusion matrices for all ensemble models

Training-test dataset (900 instances)			
M0	**Employee**	**Owner**	**Other**
	308	74	0
	31	387	0
	83	17	0
M1	**Employee**	**Owner**	**Other**
	308	74	0
	31	387	0
	83	17	0
M2	**Employee**	**Owner**	**Other**
	193	189	0
	144	274	0
	53	47	0
M3	**Employee**	**Owner**	**Other**
	308	74	0
	31	387	0
	83	17	0
M4	**Employee**	**Owner**	**Other**
	308	74	0
	31	387	0
	83	17	0
M5	**Employee**	**Owner**	**Other**
	122	260	0
	98	320	0
	36	64	0
M6	**Employee**	**Owner**	**Other**
	197	185	0
	156	262	0
	64	36	0
M7	**Employee**	**Owner**	**Other**
	150	232	0
	139	279	0
	46	54	0
M8	**Employee**	**Owner**	**Other**
	163	219	0
	142	276	0
	39	60	1
M9	**Employee**	**Owner**	**Other**
	306	76	0
	30	388	0

Validation dataset (221 instances)			
M0	**Employee**	**Owner**	**Other**
	92	12	0
	0	80	0
	32	5	0
M1	**Employee**	**Owner**	**Other**
	92	12	0
	0	80	0
	32	5	0
M2	**Employee**	**Owner**	**Other**
	46	58	0
	28	52	0
	18	19	0
M3	**Employee**	**Owner**	**Other**
	92	12	0
	0	80	0
	32	5	0
M4	**Employee**	**Owner**	**Other**
	92	12	0
	0	80	0
	32	5	0
M5	**Employee**	**Owner**	**Other**
	51	53	0
	33	47	0
	18	19	0
M6	**Employee**	**Owner**	**Other**
	62	42	0
	29	51	0
	22	15	0
M7	**Employee**	**Owner**	**Other**
	50	54	0
	25	55	0
	17	20	0
M8	**Employee**	**Owner**	**Other**
	48	56	0
	31	49	0
	19	18	0
M9	**Employee**	**Owner**	**Other**
	92	12	0
	0	80	0

	83	17	0
M10	**Employee**	**Owner**	**Other**
	191	191	191
	158	260	158
	51	49	51
M11	**Employee**	**Owner**	**Other**
	201	181	0
	146	272	0
	61	39	0

	32	5	0
M10	**Employee**	**Owner**	**Other**
	66	38	0
	44	36	0
	24	13	0
M11	**Employee**	**Owner**	**Other**
	60	44	0
	31	49	0
	23	14	0

6. Conclusion

The answers to GUESSS 2018 survey provide valuable information about students' career development preferences that can be used adapt lectures and contents to help students develop the skills needed either to start their business or to facilitate their access and adaptation to companies. When working with machine learning schemes, either single or ensemble algorithms, it is preferable to apply a cross-validation (k=10) scheme to get rid of the influence of partitioning randomness instead of simply holding out a portion of data. Averaging ten (k=10) results will be always better than having only one number. Furthermore, running replications help reduce the bias caused by the fold partitioning. Having a validation dataset is an important part of the investigation, it helps determine whether the models can generalize properly or not. The results show that half of the models proposed in the research (M0, M1, M3, M4, and M9) can predict correctly around 77% of the surveys in the training and test dataset, and in the validation dataset too. Even though, prediction ratios are similar there are significant differences in the number of attributes of each model. Finally, prediction ratios and confusion matrices suggest that the proposed classification models based on ensemble schemes do help predict mid-term professional career development preferences among university students using objective criteria based on multiple attributes.

Acknowledgements

The authors express their deep gratitude to The Global University Entrepreneurial Spirit Students' Survey (GUESSS) and to those who are involved in this important initiative.

References

Ajzen, I., The theory of planned behavior.*Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211), 1991

Bagheri, A., Pihie, Z., Factors influencing students' entrepreneurial intentions: The critical roles of personal attraction and perceived control over behavior. *The International Journal of Management Science and Information Technology (IJMSIT)*, vol. 1, no. 16, pp. 16–28, 2015

Bosma, N., Hill, S., Kelley, D., Guerrero, M., Schott, T., and Ionescu-Somers, A. GEM Global Entrepreneurship Monitor 2020/2021. *GEM Global Entrepreneurship Monitor*, 2021.

Davis, J., Goadrich, M., The Relationship Between Precision-Recall and ROC Curves, *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Divina, F., Gilson, A., Gómez-Vela, F., Garcia, M., and Torres, J., Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting, *Energies 2018*, 11, 949, 2018.

Franke, N., Lüthje, C., Entrepreneurial Intentions of Business Students - A Benchmarking Study, *International Journal of Innovation and Technology Management*, vol. 1, no. 3, pp. 269-288, 2004.

Sieger, P., Fueglistaller, U., and Zellweger, T., Entrepreneurial Intentions and Activities of Students across the World. *International Report of the GUESSS Project 2011. St.Gallen: Swiss Research Institute of Small Business and Entrepreneurship at the University of St.Gallen*, 2011.

Gorgievski, M. J., Stephan, U., Laguna, M., and Moriano, J., Predicting Entrepreneurial Career Intentions: Values and the Theory of Planned Behavior, *Journal of Career Assessment*, 26(3), 457–475, 2018.

Israr, M., Saleem, M., Entrepreneurial intentions among university students in Italy, *Journal of Global Entrepreneurship Research*, vol. 8, no. 1, 2018.

McClelland, D., Characteristics of successful entrepreneurs. *The Journal of Creative Behavior*, vol. 21, no. 3, pp. 219-233, 1987.

Powers, D., Atyabi, A., The Problem of Cross-Validation: Averaging and Bias, Repetition and Significance. *2012 Spring World Congress on Engineering and Technology*, SCET 2012 - Proceedings. 1-5, 2012.

Prada, M., Rucci, G., Instrumentos para la medición de las habilidades de la fuerza de trabajo, *Banco Interamericano de Desarrollo (Ed.), NOTA TECNICA No IDB-TN-1070*, 2016.

Prokop, D., University entrepreneurial ecosystems and spinoff companies: Configurations, developments and outcomes. *Technovation*, vol 107 (March), 2021.

Saito, T., Rehmsmeier, M., The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, *PLoS ONE 10(3): e0118432*, 2015

Witten, I., Frank, E., Hall, M., and Pal, C., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Zellweger, T., Sieger, P., and Halter, F., Should I stay or should I go? Career choice intentions of students with family business background, *Journal of Business Venturing*, vol. 26, no. 5, pp. 521–536. 2011.

Biographies

Galo Paiva is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de Santiago de Chile, Chile, and Doctor of Business Management from Universidad Autónoma de Madrid, Spain. He has taught lectures in Strategic Management, Operations Management, Industrial Engineering, and Project Planning & Management. His research interests include manufacturing process simulation, industrial design, business management, and entrepreneurship.

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, Master of Sciences in Computational Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. He has taught lectures in Discrete Event Simulation, Supply Chain Management, Engineering Economics, Corporate Finances, Financial Engineering, Business Analytics, Data Mining and Machine Learning for engineering students. He has developed a professional career working for large multinational companies (PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme). He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at TU Braunschweig, Germany. His research interests include manufacturing process simulation, transportation systems simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.