A Logistic Regression Model to Predict the Induction of Professional Baseball Players into the Hall of Fame

Jesús Adrián Betancourt Ferreiro, Daniela Chaveznava Soto, Sofía Juárez Garza and Debanie Alejandra García Bustos
Undergraduate Student
Department of Computer and Industrial Engineering
Universidad de Monterrey
San Pedro Garza García, NL 66238, México
Jesus.betancourt@udem.edu.mx, Daniela.chaveznava@udem.edu.mx,
Sofía.juarez@udem.edu.mx, Debanie.garcia@udem.edu.mx

Fernando González-Aleu
Associate Professor and Researcher
Department of Computer and Industrial Engineering
Universidad de Monterrey
San Pedro Garza García, NL 66238, México
Fernando.gonzalezaleu@udem.edu

Jesús Vázquez-Hernández
Professor, Advisor, and Researcher
Department of Master´s in Engineering Management
Instituto Tecnológico y de Estudios Superiores de Monterrey
San Pedro Garza García, NL 66238, México
jesus.vazquez@tec.mx

Abstract

In recent years, trading of sports cards has created a great attraction among collectors and investors. Although sports cards investment is considered for many a high risk, this risk could be reduced creating a long-term portfolio. For example, buying rookie cards (RC) from professional sports players in the early years of his/her career and holding them expecting until they are considered for the Hall of Fame. However, it is difficult to predict whether a player will likely be nominated for the Hall of Fame since their career depends on many factors on and off the field. Therefore, the purpose of this research is to create a model that helps collectors and investors to predict if a professional sports player will be inducted into a Hall of Fame (HOF). To achieve this aim, the authors focused their work on the Baseball game, used the SCRUM methodology, and analyzed current Baseball Hall of Famer players performance statistics. The final logistic regression model consisted of seven independent variables and predict the induction of current baseball players in the HOF with an accuracy of 98 percent.

Keywords
Baseball, Logistic Regression, Scrum, agile, prediction model

1. Introduction

Trading cards is one of the most popular collectibles around the world; including the 1886 Goodwin & Company baseball cards, the baseball cigarette cards, candy and gum set from different sports, non-sport cards (e.g., Star Wars, TV series, and other movies), game cards (e.g., Pokémon and Magic Gathering), and the current exclusive autograph and memorabilia sets (Huiges, 2018). Currently, trading cards is a multi-million-dollar industry. From January to October 2022 trading cards transactions from eBay and other auction houses were more than 900 million dollars (Vintage Card Price, 2006), attracting the attention of companies outside this industrial sector such as Fanatics (a global leader in licensed sports merchandise) which acquired Topps, an iconic baseball card company with more than 50 years in the industry (Hajducky, 2022).

The collection and exchange of sports cards is an industry continuously attracting the attention of new inventors and investors, creating new companies to assess market niches. For example, inventors developed tracking card values software and trading card suppliers (e.g., grading card protectors and stands). On the other hand, people with different economic backgrounds arrive in the trading card industry such as short-term and long-term investors. Short-term investors are those interested in obtaining a return on their investment by flipping or selling their cards from a week to a sports season, usually according to the professional sports player's short-term performance. Contrary, long-term investors hold their trading cards thinking about the professional sports players' career performance, guessing that he or she could be the next Hall of Fame (HOF) or a legend.

For long-term investors will be very useful to predict if a rookie or a sport professional player in their early career years could be a Hall of Fame player (HOFer); however, to the authors' knowledge, this industry niche is still not deeply addressed. Therefore, the purpose of this research is to create a statistical model to predict which sports players will be future HOFer. Considering sports popularity around the world, sports professional players' statistics, and the research project time scope (one month), the research team decided to limit the statistical model only to professional baseball players. To achieve this aim, the authors used a SCRUM frame or methodology, conducting several sprints to identify the sport that best suit this research project (in time and answer the research question: which are the professional baseball players' performance variables related to the induction of a HOF?

The remaining sections of this paper addressed the following topics. First, the literature review section mentioned important facts about the baseball HOF induction process and investigations related to simulated professional baseball players' performance. Second, the method section describes how the SCRUM frame was used and the data collection process. Third, the result section includes the outcome obtained from each sprint conducted, including the set of variables that predict the induction of a professional baseball player to the HOF. Lastly, in the conclusion section, the authors summarize their findings, address project limitations, and propose future research.

2. Literature Review

2.1 The National Baseball Hall of Fame

The National Baseball Hall of Fame (NBHOF) was created in 1936 with the induction of the first-class Ty Cobb, Walter Johnson, Christy Mathewson, Base Ruth and Honus Wagner (National Baseball Hall of Fame, n.d.). Today, the NBHOF has included about 340 individuals who have contributed to the growth of baseball as players (from the Major League of Baseball), managers, umpires, or pioneers/executives.

There are two processes that baseball players could follow to be inducted into the HOF: the Baseball Writers Association of America (BBWAA) and the Era Committees. First, BBWAA Screening Committee, which is formed by BBWAA members (active or honorary members), prepared the ballot list in alphabetical order of eligible candidates that check the following criteria: received a vote on a minimum of five percent of the ballots cast in the previous election or are eligible for the first time with the nomination of two of the six members of the BBWAA. BBWAA members will receive the ballot and will vote for no more than 10 eligible candidates. Their vote should be based on the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team (s) on which the player played. Any candidate receiving 75% of the ballots cast shall be elected to the membership of the NBHOF (National Baseball Hall of Fame, n.d.).

If a player was included in the ballots for 10 years and is not elected to the NBHOF, then he could be elected using the Era Committee system. The Historical Overview Committee (10-12 representatives) identify eight candidates from the Contemporary Baseball Era Players, Contemporary Baseball Era Non-Players, and Classic Baseball Era ballot. Each member of the Historical Overview Committee may vote from zero to three eligible candidates. All candidates receiving 75% or more of the ballots will earn election (National Baseball Hall of Fame, n.d.).

2.2 Prediction Models to Simulate

Some studies addressed the topic of predicting future baseball HOFers, such as Kaufman and Kaufman (1996), Findlay and Reid (2002), Cohen (2004), Mills and Salaga (2011), and Zemler (2013). First, Kaufman and Kaufman (1996) used a multiple regression analysis to predict pitchers' HOFers, founding that 60 to 73% of the variance in the pitcher's HOF selection could be explained by the predictors selected. Second, Findlay and Reid (2002) created two prediction model findings that the model that uses individual performance variables is better. Third, Cohen (2004) confirmed that most of the papers about elections in baseball HOF use regression analysis; he focused on a binary model. His research found training and testing accuracies in 99% and 97% respectively. Fourth, Mills and Salaga (2011) created a random forest algorithm for binary classification (for hitters and pitchers), obtaining 0.91% of error in the accurate forest. Lastly, Zemler (2013) created a logistical regression model to predict future hall of fame inductions by position.

To the authors' knowledge, the current literature available shows three opportunities for theoretical contributions: lack of research conducted using logistic regression, describing the individual performance variables that increase the accuracy of the model, and validation of model vs. reality. These topics will be addressed in the following sections

3. Methods

Scrum is a process through which collaborative practices are applied to obtain the best possible result of a project. It allows the execution of complex projects in a flexible manner. It consists of regular partial deliveries of the finished product (Sutherland and Sutherland, 2021). Due to its collaborative nature Scrum must consist of a Scrum Master, a Product Owner and the Scrum team (Srivastava et al., 2017).

The Product Owner is responsible for adding value to the final product delivered by the scrum team. Product owners are accountable for understanding the needs of the customers or otherwise, the product will not satisfy the customer's needs. For this work, the product owner was Universidad de Monterrey (referred to as UDEM). Scrum Masters have the goal of guiding the Scrum team through their knowledge in the methodology and are responsible for the proper follow-up of it. Scrum Masters are also responsible for ensuring that the daily meetings are fulfilled in order to achieve the project objectives. For this work, the Scrum Master was the Senior Capstone Project advisor. Finally, the Scrum Team is responsible for the development and execution of the product (see Figure 1).

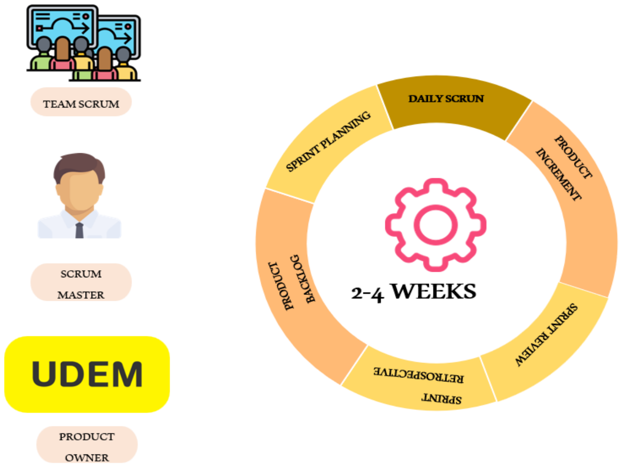


Figure 1. SCRUM Framework

This methodology is composed of *sprints*, these are a series of 1 to 3 weeks blocks of work with assigned tasks. The goal of the sprint is decided first through a *Sprint Backlog* in which the product owner enlists the requirements of the sprint. Once the sprint has started, the scrum team participates in another key activity of the methodology, the *Daily Scrum*. The daily scrum is a 15-minute-long meeting where every team member is involved. When the sprint has finished, the scrum team alongside the product owner do carry out the *Sprint Revision* to understand the progress. Sprint revision is an adaptation and inspection activity, it allows the product owner to see the progress of both the team and the product. The *Sprint Retrospective* is done after the Review and involves inspecting and adapting the process. The retrospective serves to provide an opportunity for effective communication from the whole team about what is working and what is not. In addition to agreeing on the changes to be made (Deemer et al., 2009).

In order to carry out the creation of a model to predict the induction of future professional baseball players to the HoF was performed using R Studio. We chose this software because it is a processing program which allows the download of a wide range of tools and therefore the scope of learning is greater. It also has great management and effective data storage, a wide collection of tools for data analysis, its way of visualizing data through the graphs provided by the software, a friendly programming language that the entire Scrum team has the ability to handle, the creation of automatic reports and high-quality data visualizations.

During the first sprint, we first conducted thorough research on the HOF, to know the variables involved, to analyze statistics on the inducted players in order to create the hypotheses and objectives to work for this project. At the end of the data analysis, the database found in R Studio was emptied to run the mathematical model proposed with two important variables (number of Hits and number of Home Runs).

Moving on to the second sprint, we ought to compare our model with the real data in order to make adjustments if necessary. Once these adjustments have been made, we will proceed to increase the number of variables to be used in the proposed model. Finalizing with the interpretation of the data provided by the software.

For the third sprint, the number of variables in the model will be increased again for better delivery of results. In addition, it is desired to represent the players who enter the HoF through a graph for better visualization and understanding. Once finished, the hypotheses raised from the beginning of the research will be tested to end with our conclusions.

Our last sprint will be concluded with the delivery of the official documents with all the information collected from the beginning.

4. Data Collection

Prior to the formulation of the mathematical model, three initial constraints were identified for the realization of the model. We decided not to include players on baseball's ineligible list because players with a bad reputation will affect the expected results. On the other hand, we did not take pitchers into account since this would result in a completely different model than the one desired. Finally, it was established to test the model with all HOF members elected by the BBWAA as this research is focused only on Baseball Writers of America and not on the Veterans Committee, since they can only choose players who can no longer participate through the BBWAA. Table 1 shows the variables that were taken into account initially.

Table 1. Potential variables to use in the model with their type and definitions

Variable	Variable Type	Definition
HoF	Binary Response	Indicator of whether a player is in the Hall of Fame (denoted by a 1) or not in the Hall of Fame (denoted by a 0)
HG	Continuous	Hits/played games
HRG	Continuous	Home runs/played games
AB	Continuous	At Bats
R	Continuous	Runs; the number of times a player scored a run by crossing home plate in their career.
RBI	Continuous	Runs Batted In; the career number of times a player’s offensive actions led to another player scoring a run by crossing home plate.
SB	Continuous	Stolen Bases

BB	Continuous	Bases on Balls/Walks
BA	Continuous	Hits/At Bats
OBP	Continuous	On Base Percentage; (H + BB + HBP)/(At Bats + BB + HBP + SF)
SLG	Continuous	Total Bases/At Bats or (1B + 2*2B + 3*3B + 4*HR)/AB
OPS	Continuous	On-Base + Slugging Percentages
HRAB	Continuous	Home Runs/At Bats
RBIG	Continuous	Runs Batted In/played games
SBG	Continuous	Stolen Bases/played games
BBG	Continuous	Bases on Balls/played games
MVP	Discrete	The number of Most Valuable Player awards a player won
RoY	Discrete	Rookie of the Year Award
ASGA	Discrete	All-Star Game Appearances
GG	Discrete	Golden Glove Award
BC	Discrete	Batting Champion

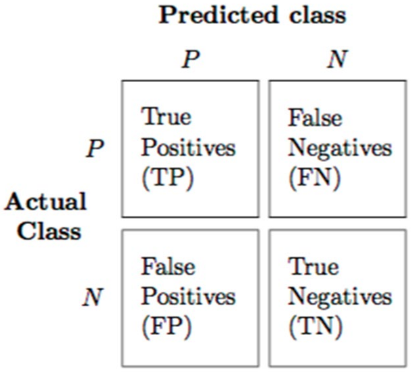
For better progress in the delivery, trial-error models were performed to discard variables that do not add value to the model and add variables that added value. Table 2 shows the variables that were used for each Sprint. For the first Sprint, Hits and Home Runs per game played were added according to the statistics presented by Springer (2021), which we considered the most relevant to test first. At the end of the first model, we realized that there were more variables that could add value and increase the accuracy of the model.

Table 2. Variables per Sprint

Sprint	**Variable**
1	HG and HRG
2	AB, R, RBI, SB, BB, BA, OBP, SLG, OPS, and HRAB
3	HRAB, SBG, BBG, BA, OPS, MVP, ROY, ASGA, GG, and BC

Now, for the classification of the model a confusion matrix was made in which different metrics were used, which are shown below (see Figure 2):

- True Positives (TP): Model correctly predicts that the player has entered HoF.
- False Negatives (FN): Model incorrectly predicts that the player has not entered HoF.
- False Positives (FP): Model incorrectly predicts that the player has entered HoF.
- True Negatives (TN): Model correctly predicts that the player has not entered HoF.



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2. Shows the nomenclature of each of its quadrants, where:

As with any statistical model, we established the metrics to evaluate the model and the expected value for each metric (see Table 3 and Table 4). The expected value criterion is a model based on the probabilities of occurrence of a particular event and results from the weighted sum of the payments corresponding to the decision alternative.

Table 3. Metrics

Sprint	Variable	Importance
Accuracy	Number of positive predictions that were correct.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Percentage of positive cases detected.	$\frac{TP}{(TP+FP)}$
F1 Score	Summarizes accuracy and sensitivity in a single metric.	$\frac{(2)(Precision)(Sensitivity)}{Precision+Sensitivity}$
Sensitivity	It is the proportion of positive cases that were correctly identified by the algorithm.	$\frac{TP}{(TP+FN)}$
Specificity	These are negative cases that the algorithm has correctly classified	$\frac{TN}{(TN+FP)}$

Table 4. Expected Value for each Metric

Metric	Bibliographic reference results	Expected results of our model
Accuracy	0.8300	0.8500
Precision	0.8300	0.8500
F1 Score	0.8000	0.8500
Sensitivity	0.83000	0.8500
Specificity	0.8000	0.8500

5. Results and Discussion

The 220 players as well as the selected variables were registered into the R Studio software, which helped to visualize the information required for the model to work. The results are grouped by general results and hypotheses raised at the beginning. Find the expected value of the precision and accuracy of the regression model. As mentioned above, the project was divided into 4 sprints and within each sprint, a result was achieved.

5.1 Sprint 1

As a first step, we established the research question and set the research hypotheses. On the other hand, in order to obtain other models with which to compare ours, we resorted to an exhaustive investigation of articles and journals based on logistic regression. We identified descriptive and explanatory variables relevant to the model, as well as identified the relationships between variables based on literary theory. We documented which players meet the requirements to enter the HOF, defined restrictions, formulated the mathematical model and tested the model with 2 variables: Hits and Home Runs. These variables were chosen as the first test, because they are the most important achievements of a player in the game.

During this sprint we realized there was a bias in our result because there were neither true nor false negative cases. This was due to the fact that there were more HOFers than non HOFers in our data frame. This meant that our data selection must be adjusted and more varia. This should be added to increase the accuracy (see Table 5)

Table 5. Metric results for sprint 1

Metric	**Result**
Accuracy	0.8190
Precision	0.8190
F1 Score	0.9005
Sensitivity	1.0000
Specificity	0.0000

5.2 Sprint 2

Within this sprint, we made comparisons with real data (players already in HoF, to show that the model is working and does not need to be adjusted). And we arrived at the result presented in the previous sprint, where we adjusted and added new variables.

An important criterion to consider is the multicollinearity of our variables. Multicollinearity is the high correlation between more than two explanatory variables. The solution for severe multicollinearity was to eliminate the regressors of variables with a high linear relationship.

For this reason, we had to eliminate variables that had this multicollinearity and affected the accuracy of our model. We also established the conditions that a logistic model has.

For this sprint the accuracy of the model increased to 0.8750 resulting in a fulfillment of the expected value of the model but, the other metrics still have not reached the expected value, such as the precision that modifications need to be made to crease it to at least 0.8500 (see Table 6). There are some issues with the fairness of the model because actual players will be affected by not having the same amount of games as the retired players.

Table 6. Metric results for sprint 2

Metric	**Result**

Accuracy	0.8750
Precision	0.8000
F1 Score	0.8889
Sensitivity	1.0000
Specificity	0.7500

5.3 Sprint 3

We decided to add the awards that players can win. We made adjustments to variables by dividing them over the total games played to have predictions of players who are currently playing to be more fair.

Comparing the variables that our model marked with the highest importance, MVP andASGA coincide as the most relevant in the Aron Springer article.Higher values indicate more importance, and these results match very well with the p-values of the model. Results from Sprint 3 shows (see Table 7) that all five metrics achieve the expected value (see Table 4)

Table 7. Metric results for sprint 3

Metric	**Result**
Accuracy	0.9091
Precision	0.9500
F1 Score	0.9048
Sensitivity	0.8636
Specificity	0.9545

5.4 Sprint 4

After the fourth and final sprint, we can confidently conclude that our model is capable of predicting the inductance of players into the HoF. Table 8 shows a list of 26 players that are either retired or currently playing in MLB. Our model predicted them to enter HOFer in the near future.

Table 8. Predicted players to enter HoF

Name	**Status**
Mike Trout	Currently playing
Bryce Harpert	Currently playing
Mookie Betts	Currently playing
Luis Arraez	Currently playing
José Altuve	Currently playing
Miguel Cabrera	Currently playing

Yadier Molina	Currently playing
Ryan Braun	Retired
Howie Kendrick	Retired
Nick Markakis	Retired
Daniel Murphy	Retired
Dustin Pedroia	Retired
Hanley Ramírez	Retired
Martín Prado	Retired
Carlos González	Retired
Melky Cabrera	Retired
Ichiro Suzuki	Retired
Joe Mauer	Retired
David Wright	Retired
Matt Holliday	Retired
Victor Martínez	Retired
Brandon Philips	Retired
Denard Span	Retired
Chase Headley	Retired
Carlos Beltrán	Retired
Stehpen Drew	Retired

6. Conclusion

While progressing through the sprints, the metrics used got closer to the expected results of the model. When comparing it to the other current models we realized that our model surpassed their final results (see Table 9). Compared to the model of Springer we used fewer variables and got better accuracy. One important finding was not to use the Silver Slugger award, contrary to Springer's model because this award was only given after 1980. One key factor that helped our model to be used to predict player inductance more accurately was to divide variables such as SB, RBI and BB by the number of games played, all of these including HRAB being new variables that were not present in other models.

Table 9. Final metric results

Metric	Bibliographic reference results	Expected results of our model	Obtained results
Accuracy	0.8300	0.8500	0.9091
Precision	0.8300	0.8500	0.9500
F1 Score	0.8000	0.8500	0.9048
Sensitivity	0.83000	0.8500	0.8636
Specificity	0.8000	0.8500	0.9545

As mentioned above, the model takes a random sample for training data and the rest for validation data, so several tests were carried out with these random results and these were the maximum values that we obtained an accuracy of 90%, a precision of 95 %, F1 score of 90%, sensitivity of 86%, and specificity of 95%. Collectors and investors could use this model to improve the value of their collections or assess identifying potential Baseball HOFers in the early stages or their careers, when their Baseball cards and autos still are in reasonable prices.

In this model we found certain limitations. Among them we suppose that there must be linearity between the dependent and independent variables, it should be verified that our independent variables explain whether to enter the HoF.

Another one is to have null multicollinearity between independent variables. Also, you must manually add the new players you want to predict and perform certain calculations for some variables since such information is not found. This is the case as for the variables SBG, BBG, and HRAB . And lastly, it can only be used to predict discrete functions.

For future investigations, it is recommended to explore participation in the world series as a variable to predict the inductance of a player to the HoF. Another opportunity is to standardize the variables to make more accurate predictions for actual players, specifically for the awards received, it would likely be useful to predict how many awards the player will receive. There is a big area of opportunity regarding the creation of a model to predict if a pitcher will be inducted into the HoF, as well as exploring players that have been pitchers and batters such as Ohtani or Babe Ruth and analyze if their likelihood of entering the HoF is greater being a batter or a pitcher or if there is any way to create a model that considers both statistics (batting and pitching).

References

Cohen, D.A., EA-lect: An evolutionary algorithm for constructing logical rules to predict election into Cooperstown, *Proceedings of the 2004 Congress on Evolutionary Computation*, vol. 2, pp.1354-1361, 2004

Findlay, D.W. and Reid, C.E., A comparison of two voting models to forecast election into The National Baseball Hall of Fame, *Managerial and Decision Economics,* vol.23 no.3, pp. 99-113, 2002

Hajducky, D. Fanatics Acquires Topps´ Trading Cards and Collectables Business. *ESPN*, January 3, 2022

Huiges, B. From the Hobby to Investment. The History and Rise of Trading Cards as a Tradable Asset Class, *Financial History*, Summer 2018, pp.. 32-35, 2018.

Sutherland, J., and Sutherland, J.J., Scrum. El Arte de Hacer el Doble de Trabajo en la Mitad de Tiempo, 6th Edition, Editorial Ocenao, Estado de Mexico, 2021

Kaufman, A.S., and Kaufman, J.C., Multiple regression analysis and baseball hall of fame membership: Part II. Focus on pitchers, *Perceptual and Motor Skills*, vol.84 no.3, pp. 883-889, 1996

Mills, B.M. and Salaga, S., Using tree ensembles to analyze national baseball hall of fame voting patterns: An application to discrimination in BBWAA voting, *Journal of Quantitative Analysis in Sports,* vol. 7 no.4, 2011

National Baseball Hall of Fame. *History of the Museum.* Retrieved October 10, 2022, from https://baseballhall.org/about-the-hall, n.d.

National Baseball Hall of Fame. *BBWAA Election Rules.* Retrieved October 10, 2022, from https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election, n.d.

National Baseball Hall of Fame. *ERA Committees.* Retrieved October 10, 2022, from https://baseballhall.org/hall-of-famers/rules/eras-committees#:~:text=The%20Era%20Committees%2C%20formerly%20known,in%20one%20of%20four%20eras, n.d.

Vintage Card Price. *Market Report*. Retrieved October 6, 2022, from https://vintagecardprices.com/hot/free-baseball-card-prices, 2006

Zemler, K., A Statistical Approach to Predict future Members of the Baseball Hall of Fame, *Master Degree Thesis Unpublished,* 2013

Biographies

Jesús Adrián Betancourt Ferreiro is an Industrial and System Engineering Senior student at Universidad de Monterrey. Currently, he is working at Xpertal, a FEMSA business unit, in the Change Management Department.

Daniela Chaveznava Soto is an Industrial and System Engineering Senior student at Universidad de Monterrey. Currently, she is certified as ISO 9001:2015 internal auditor.

Sofía Juárez Garza is an Industrial and System Engineering Senior student at Universidad de Monterrey. She is a certified ISO:2015 internal auditor. Currently, she is working at Home Depot in the Operations Department

Debanie Alejandra García Bustos is an Industrial and System Engineering Senior student at Universidad de Monterrey. She is a certified ISO:2015 internal auditor. Currently, she is working at FEMSA in the Price Strategy Department

Fernando González Aleu received a BS in Mechanical and Management Engineering at UDEM, an MS at ITESM in 1999, and both an MS and Ph.D. in Industrial and Systems Engineering from Virginia Tech in 2015 and 2016, respectively. His research is focused on continuous improvement projects and performance excellence models with more than 42 publications, including books, book chapters, journal papers, and conference proceedings. He has more than 15 years of experience in higher education organizations as a faculty and ISE Undergraduate Program Director. Prior industry experience includes 15 years of implementing quality systems, environmental systems, and management systems in Mexico and Chile. He is a member of the IISE, ASEM, ASQ, and IEOM.

Jesús Vázquez Hernández is an Industrial and Systems Engineer, with a Master's Degree in Business Administration and postgraduate studies in Prospective and Strategic Intelligence, as well as a Doctorate in Strategic Planning and Technology Management. He collaborates as a Master Level Instructor at APICS and the Project Management Institute. He has more than 20 years of experience in application projects in the industry with the TLS methodology (ToC + Lean + SixSigma) and research on Industry 4.0 in the Supply Chain. He has held management positions in companies such as Vitro, Ingersoll Rand, Owens-Illinois and has been a graduate and undergraduate professor in engineering and business careers from 2008 to date at various private universities such as Tecnológico de Monterrey, Ibero and UdeM. He is currently an independent consultant and advisor in various companies.