Applying the CRISP-DM Framework for Teaching Business Analytics

By

Sanjiv Jaggia[1], Alison Kelly[2], Kevin Lertwachara[1], and Leida Chen[1]

Forthcoming at DSJIE; not for redistribution.

[1] Professor, Orfalea College of Business, California Polytechnic State University, San Luis Obispo, CA 93407; email: Sanjiv Jaggia sjaggia@calpoly.edu, Kevin Lertwachara klertwac@calpoly.edu, Leida Chen lchen24@calpoly.edu

[2] Professor, Suffolk University, 73 Tremont St., 10th Floor, Boston, MA 02108; email: Alison Kelly <akelly@suffolk.edu>

Applying the CRISP-DM Framework for Teaching Business Analytics

Abstract

Experiential learning opportunities have been proven effective in teaching applied and complex subjects such as business analytics. Current business analytics pedagogy tends to focus heavily on the modeling phase with students often lacking a comprehensive understanding of the entire analytics process, including dealing with real life data that are not necessarily 'clean' and/or small. Similarly, the emphasis on analytical rigor often comes at the expense of storytelling, which is among the most important aspects of business analytics. In this paper, we demonstrate how the philosophy of the Cross Industry Standard Process for Data Mining (CRISP-DM) framework can be infused into the teaching of business analytics through a term-long project that simulates the real-world analytics process. The project focuses on problem formulation, data wrangling, modeling, performance evaluation, and storytelling, using real data and the programming language R for illustration. We also discuss the pedagogical theories and techniques involved in the application of the CRISP-DM framework. Finally, we document how the CRISP-DM framework has proved to be effective in helping students navigate through complex analytics issues by offering a structured approach to solving real-world problems.

INTRODUCTION

The importance of incorporating business analytics in pedagogy has been well documented (see, for example, Asamoah et al., 2017 and Henke et al., 2016). This trend is further evidenced by the proliferation of business analytics courses and programs across universities and by the increasing industry demand for analytics professionals. Although there are no universally-agreed upon definitions of the term ‘business analytics’, we follow the lead supplied by the Institute for Operations Research and the Management Sciences (INFORMS) to define the term as “the scientific process of transforming data into insights for the purpose of making better decisions” (2019). Other scholars have also provided relevant interpretations of what the analytics process entails. For example, Wilder and Ozgur (2015) define business analytics as “the application of processes and techniques that transform raw data into meaningful information to improve [business] decision making.” Many business enterprises now describe themselves as “analytics-based firms” and have become heavily dependent on data-driven decision making to improve their organizational performance (Watson, 2013). As a result, business analysts in these organizations are expected to be fully knowledgeable and well-versed in analytics concepts and techniques.

The CRISP-DM (**CR**oss Industry **S**tandard **P**rocess for **D**ata **M**ining) framework is widely regarded as the most relevant and comprehensive guiding principle for carrying out analytics projects (Abbasi et al., 2016). In introducing the CRISP-DM framework, Wirth and Hipp (2000) describe business analytics as a creative process that requires a standard approach to “help translate business problems into data [analysis] tasks, suggest appropriate data transformations and data [analysis] techniques, and provide means for evaluating the effectiveness of the results and documenting the experience.” This philosophy has been adopted by business analysts and practitioners in many industry segments, regardless of the analysis techniques or computing technologies used in the project.

Despite the wide acceptance and adoption of the CRISP-DM framework by practitioners, current business analytics pedagogy fails to provide a holistic approach to analytics. Students often find themselves not adequately trained to deal with real life projects and the emphasis on analytical rigor often comes at the expense of storytelling, which is among the most important aspects of business analytics (Dykes, 2016). As suggested by Heim et al. (2005), student learning in technology-based disciplines such as business analytics can be enhanced through experiential projects that simulate real-life activities. The current teaching brief infuses the philosophy of the CRISP-DM framework into the teaching of business analytics through a term-long project that simulates the analytics process. Using real-life data that are not necessarily ‘clean’ and/or small, we demonstrate how instructors can apply the six phases of the CRISP-DM framework in hands-on analytics activities.

Benefits of experiential learning pedagogy have been well documented. For example, Burch et al. (2019) examine over 89 research studies over a 43-year span and show that students experience “superior learning outcomes” when experiential learning is used. Moreover, as an indicator of its efficacy in enhancing learning outcomes, experiential learning has been widely implemented at many universities (see, for example, Cardozo et al. (2002) and Silvester et al. (2002)). For the application in this paper, we use the programming language R for illustration, but all of the analytics tasks can be similarly completed with any other software such as Python or Analytic Solver (formerly called XLMiner). As business analytics is considered a creative process, we argue that the pedagogical focus for teaching this subject should be placed on problem formulation, data wrangling, modeling, performance evaluation, and storytelling.

THE CRISP-DM FRAMEWORK

CRISP-DM was developed in the 1990s by a group of five companies: SPSS, TeraData, Daimler AG, NCR, and OHRA (Wirth & Hipp, 2000). CRISP-DM consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The six phases can be summarized as follows:

- ✓ Business understanding: According to Wirth & Hipp (2000), this first phase focuses on "understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives." Students should be reminded that this is a critical step where business objectives are identified in order to steer the subsequent direction of the project.
- ✓ Data understanding: This phase involves an initial data collection and proceeds to activities that help students become more familiar with the data. Students also need to identify potential data quality problems, preliminary insights into the data, and possible subsets of data to form hypotheses that can uncover hidden information.
- ✓ Data preparation: Specific tasks in this phase include data reduction, data wrangling and cleansing, and data transformation (e.g., creating dummy variables) for subsequent analyses and testing.
- ✓ Modeling: This phase involves the selection and development of analytics techniques and models. In addition, portions of a data set are often set aside for training and validating the model(s).
- ✓ Evaluation: This phase involves reviewing and interpreting the analysis results in the context of the business objectives and success criteria described in the first phase.

- ✓ Deployment: During this final phase, the knowledge gained from data analysis is translated into a set of actionable recommendations. In addition to performing appropriate analysis, analysts need to understand that effectively communicating the analysis results to business constituents also plays a key role in a successful analytics project.

The CRISP-DM framework implies a cyclical nature of business analytics projects, and therefore, is often depicted as a life cycle model as shown in Figure 1.

Insert Figure 1 Here

In addition to being implemented in industry projects, the CRISP-DM framework has also been used as a guiding principle in curriculum development in higher education. For example, in 2018, the University of Chicago launched a data analytics program whose core courses are "structured along the CRISP-DM methodology (2019)". Other programs in analytics (see, for example, Northwestern University's Master's in Data Science program, 2019) often incorporate the CRISP-DM methodology in their curriculum. However, for individual analytics courses, the pedagogical focus is usually on the modeling phase, which is only one of the six phases in the CRISP-DM framework (Rudin, 2012). Students often fail to realize the interplays between the phases and how they collectively contribute to the success of analytics projects. Moreover, most data sets introduced in these courses tend to be small and 'clean' with little, if any, data wrangling required. As such, the data understanding phase and the data preparation phase receive minimal coverage. Finally, as a technical field, analytics education tends to overlook the importance of storytelling, the process of translating data points, analytical methodologies, and findings into interesting stories that are more palatable to the intended audience. The project described in this teaching brief aims to address these deficiencies witnessed in the current analytics education.

THE PROJECT

The project can be assigned to teams of students enrolled in a business analytics course in either the upper division undergraduate curriculum or at the graduate level. Unlike the pedagogical approach used by Anderson and Williams (2019) where students work on personal analytics topics such as learning foreign languages, stress management, and personal wellbeing, the current research focuses on the business processes integrated in the CRISP-DM framework. Our approach offers students a culminating experience that involves a practical business topic with project requirements that permeate throughout the CRISP-DM process. Students are advised to work on the project throughout the term as each of the six CRISP-DM phases is discussed. The instructor will conduct periodic reviews of the project progress and provide feedback to student teams. We provide a sample of the project assignment in Appendix A.

The learning objectives of this project align with the CRISP-DM framework as shown in Table 1. In the first phase, students formulate business questions that will ultimately lead to business strategies or actions. They then describe the data in terms of the business context, and perform data wrangling to prepare the data for subsequent analyses. Following data description and data wrangling, students apply modeling techniques to the final data set(s) to produce a number of predictive models that best address the business questions. They then evaluate model performance to select the best predictive model(s). Finally, as the deployment of the predictive model(s) is not practical in an educational setting, students focus on communicating key findings of the project through storytelling in the last phase.

Insert Table 1 Here

The Data

The data used in this project is the ERIM data set provided by the James M. Kilts Center of the University of Chicago's Booth School of Business (https://www.chicagobooth.edu/research/kilts/datasets/erim). The ERIM data set contains demographic information on 3,189 households in two midwestern cities in the United States and their purchases in several product categories (e.g. frozen dinners, yogurt, ketchup, margarine, etc.) from participating stores over a three-year period (1985 – 1988). For demonstration purposes, the application used in this paper focuses only on yogurt and frozen dinners. The instructor may determine the scope of the project that best aligns with the objectives of the course. One approach is to ask each team to select a product category to analyze, and another approach is to design the project as a competition where student teams all focus on the same project category or categories.

It is worth noting that while the project described in this teaching brief was designed around an existing data set, real life business analytics projects would likely start with business managers identifying problems that require data-driven solutions instead of asking what questions they can answer with the existing data. The students must understand that the identified business questions should drive the entire analytics process – including the acquisition of relevant data. The instructor should explain this important limitation of the project to the students in order to provide them with a realistic expectation of what they are likely to encounter in real projects.

Business Understanding

The first phase of the project deals with formulating business questions through an understanding of the business context. As most students are familiar with the retail industry, students can identify the potential business opportunities the data set presents to retailers and manufacturers in marketing and selling these products. For consistency across student teams, we suggest that the following two business questions be included in the assignment:

1. Which households are likely to purchase yogurt and frozen dinner products?
2. How much money is each household likely to spend on each product?

Students also need to understand that in practice, documenting available resources, estimating potential costs and benefits of the project, and identifying success criteria are also part of this first phase. In addition, business analysts often work with a wide range of stakeholders, and therefore, identifying relevant stakeholders and understanding their values and requirements are critical to the success of the project. During this phase of the project, the instructor may also impress upon students the general differences between supervised and unsupervised techniques. For example, students may be asked to consider whether supervised (predictive modeling) or unsupervised (pattern recognition) learning would be appropriate for achieving the analytical objectives and whether the current data set supports these techniques. These questions can be further explored during the data understanding phase of the CRISP-DM framework.

Data Understanding

Depending on the prerequisite knowledge of the students, the instructor can choose to require students to download the original data from the ERIM website, which require a fair amount of data integration and pre-processing. The original household data set contains 3,189 observations with 62 variables. Given that we are interested in variables that may influence a household's decision to purchase yogurt or frozen dinners, we remove irrelevant variables from the data set such as the head of household's first name, whether the household had a washer and dryer, etc. Also, because nearly all male and female households are white in this data set, we delete the race variables. After integrating the household data with detailed purchase records, we produce the modified data set, now renamed ERIMData, which contains 3,189 observations and 18 variables. We provide a complete list and description of the variables in Table B1 in Appendix B; all of the data used in this application are available upon request.

Even with a more manageable data set, data preparation and wrangling are still necessary prior to model development and analysis. In this application, data wrangling and analytical modeling are performed using the R language; however, these are common tasks that can also be completed with other software packages or programming languages. For those who wish to learn R, basic tutorials can be found at www.r-project.org/about.html and www.rstudio.com/online-learning. During this phase, students also explore the data and identify possible variables that may add value to subsequent analysis phases. In Appendix C, we provide a portion of the R code used for data wrangling and selected business analytic models. The complete R code is available upon request.

It is a common practice to produce summary statistics, look for symmetry, and/or identify outliers for key variables. Table 2 provides a summary for the two expenditure (target) variables. Even with only descriptive statistics, students can draw insights from the data. For both target variables, the median is notably less than the mean and the maximum value is dramatically higher than the third quartile. Thus, it is likely that both distributions are positively skewed and have outliers. Students are encouraged to use data visualization, such as boxplots, to reinforce this finding, and to explore other visualization tools, such as histograms, stacked column charts, scatterplots, bubble plots, and heat maps to discover other interesting patterns and stories.

Insert Table 2 Here

The strong evidence of positive skewness and/or outliers suggests the following two approaches for conducting predictive analytics:

1. Log-transform the yogurt and dinner expenditure variables for prediction models.
2. Bin the yogurt and dinner expenditure variables for classification models.

These transformations along with other potential data-related issues can then be dealt with in the data preparation phase of the CRISP-DM framework.

Data Preparation

Although the ERIM data set is from the 1980's, it highlights so many of the relevant issues that business analysts still confront on a daily basis, starting with transforming unwieldy raw data into useful, actionable information. During the data preparation stage, we ask students to develop a comprehensive plan for data wrangling, which is the process of cleansing, integrating, transforming, and enriching the data. Most analysts will attest that data wrangling is one of the most critical and time-consuming steps in any analytics project.

With 3,189 observations and 18 variables in ERIMData and preliminary ideas for potential analysis techniques, students are advised that further data preparation is still necessary. In many data sets, including this one, there are missing values. For this project, it is appropriate to replace all missing values with 0's. For example, missing information on male head of household implies that there is no male head of household. Similarly, missing information on yogurt expenditure implies that there is no expenditure on yogurt. The instructor, however, should remind students that it is not always appropriate to replace missing values with 0's.

Next, students are told that data wrangling is needed to create/transform variables that make good economic sense for the analysis. We create 12 additional variables in this phase. For instance, we create a Yogurt variable that assumes the value of 1 if a household purchases yogurt over the time period, 0 otherwise. Similarly, we create a Dinner variable that assumes the value of 1 if a household purchases frozen dinners over the time period, 0 otherwise. With the 12 new variables, we no longer need 13 of the original variables, and therefore, we remove them to create the final data set for analysis with 3,189 observations and 17 variables (18 + 12 – 13 = 17). We provide a complete list and description of the newly created variables as well as the ones retained from the original data set in Table B2 of Appendix B.

It is always useful to produce descriptive statistics on key variables in the final data set with a focus on the business questions previously discussed. One possible project requirement is to have students subset the data based on whether or not the households purchased the two products of interest: yogurt and frozen dinners. Table 3 shows averages for potential predictor variables for the entire sample and for subsetted households.

Insert Table 3 Here

Students are encouraged to grasp key characteristics of the data. They can also draw some interesting observations from subsetted data. For example, on average, households that purchased yogurt earned more, worked longer hours, and had a younger head of household as compared to households who did not purchase yogurt. Similar observations can be drawn when comparing households that did and did not purchase frozen dinners.

Modeling

The modeling process involves applying predictive models to the data to identify hidden structures, patterns and relationships among variables. The efficacy of competing models is assessed in this phase as well. Therefore, we tend to divide this phase into two sub-phases, model development and model assessment.

Model Development

In the model development sub-phase, the instructor should highlight the strengths and limitations of various modeling techniques. Students will identify and choose the appropriate analytical techniques after considering their advantages and limitations. Initially, students may start with a model that offers a higher level of interpretability. For example, both logistic regression for classification and multiple linear regression for prediction are highly interpretable and quantify the impact of each predictor variable on the target variable.

Table 4 shows the logistic regression results for classifying whether or not a household purchases yogurt and frozen dinner products, respectively. As part of the project assignments, the instructor should ask students to consider the following questions based on the initial modeling results: (a) which predictor variables are the most influential predictors?, (b) how much impact does each predictor variable have on the probability of a household purchasing yogurt/dinner?, and (c) which type of household is likely to purchase yogurt/dinner? The instructor may also ask students to compare answers to these questions to their initial assumptions gained from data exploration during the earlier phases.

Insert Table 4 Here

Students are also asked to perform multiple linear regression models on the expenditure variables. As mentioned before, given the skewness of the numerical target variables, these target variables must first be transformed into natural logs. The instructor should provide the correct interpretation of the estimated coefficients of log-linear models. Qualitatively, the results from the multiple linear regression models are similar to those of the logistic models; the results are not reported in the paper for the sake of brevity.

While logistic and linear regression models offer important insights on how the predictor variables influence the target variables, the instructor can remind students that predictive modeling focuses more on the model's ability to classify or predict a future case correctly than trying to interpret or draw inferences from the model. In other words, a well performing explanatory model may not necessarily be a good predictive model. Data-driven techniques such as naïve Bayes, ensemble trees and *k*-nearest neighbors may result in better predictive models even though they suffer in interpretability.

Model Assessment

In the model assessment sub-phase, the instructor should stress the importance of evaluating model performance using the validation or test data set instead of the training set. Performance measures should evaluate how well an estimated model will perform in an unseen sample, rather than making the evaluation solely on the basis of the sample data used to build the model. The validation data set not only provides measures for evaluating model performance in an unbiased manner but also helps optimize the complexity of predictive models.

Students are asked to evaluate model performance using the validation data set each time a model is developed and focus on measures of predictive performance rather than on goodness-of-fit statistics as in a traditional analytical process. For classification models, performance measures include the accuracy rate, sensitivity and specificity. For prediction models, common performance measures are the mean error (ME), the root mean square error (RMSE), and the mean absolute error (MAE). Furthermore, performance charts such as the cumulative lift chart, the decile-wise lift chart, and the receiver operating characteristic (ROC) curve are also used to evaluate model performance.

To illustrate the teaching points, we partitioned the data to re-estimate and assess the logistic regression model for classifying whether a household would purchase yogurt or frozen dinners, respectively; see Table 5 for the performance measures. We present two sets of performance measures, using the cutoff value of 0.5 (the default cutoff value for binary classification models) and the cutoff value equal to 0.82 for yogurt and 0.33 for frozen dinners (the actual proportion of households in the data that purchased yogurt and frozen dinners, respectively). It is important to point out to students that classification performance measures are highly sensitive to the cutoff values used. A higher cutoff value classifies fewer number of cases into the target class, whereas a lower cutoff value classifies more cases into the target class. As a result, the choice of the cutoff value can influence the confusion matrix and the resulting performance measures. In cases where there are asymmetric misclassification costs or an uneven class distribution in the data, it is recommended that the proportion of target class cases be used as the cutoff value. For example, by setting the cutoff value to 0.33 for frozen dinners, the model generates a sensitivity value of 0.5671 meaning that 56.71% of the target class cases are correctly classified, versus a sensitivity value of 0.1106 if the cutoff value is 0.5.

Insert Table 5 Here

It is sometimes more informative to have graphical representations to assess model performance. Figure 2 displays these charts that are associated with the logistic regression model for frozen dinners. Note that the charts are created using the validation data set. Unlike the numeric performance measures, the performance charts are not sensitive to the choice of cutoff value. Students need to be able to articulate the performance of various models based on the performance charts. For example, Figure 2 suggests that the logistic regression model offers improvement in prediction accuracy over the baseline model (random classifier). The lift curve lies above the diagonal line suggesting that the model is able to identify a larger percentage of target class cases (households that purchase frozen dinners) by looking at a smaller percentage of the validation cases with the highest predicted probabilities of belonging to the target class. The decile-wise lift chart conveys similar information but presents the information in 10 equal-sized intervals. Finally, the ROC curve also suggests that the model performs better than the baseline model in terms of sensitivity and specificity across all possible cutoff values. The area under the curve (AUC) value is 0.6138, which is larger than the AUC value of the baseline model (AUC = 0.5).

Insert Figure 2 Here

In the modeling phase of the CRISP-DM framework, students were asked to consider classification and prediction models. We show the logistic regression model for classification and the multiple linear regression model for prediction. While the logistic model seemed to work well for both products in this application, students should verify that other data-driven techniques, such as ensemble trees and *k*-nearest neighbors, do not yield better performance measures. As an alternative to the multiple linear regression model, students may want to consider a regression tree. Figure 3 displays the regression tree model and its performance measures for frozen dinners, which can be used as the basis for model selection. For the regression tree, the amount a household spends on frozen dinners is determined by the number of members in the household and age. Based on RMSE, this model has a smaller average prediction error than the multiple linear regression model. Students are encouraged to verify this fact and contemplate why a simpler model may produce more accurate predictions than the more complex ones do in some cases. While we focus on supervised learning in this application, students can also explore unsupervised learning approaches (e.g. cluster analysis) to further examine interesting patterns and relationships that may exist in the data.

Insert Figure 3 Here

Evaluation

While the efficacy of the predictive models with regard to various performance measures is assessed during the modeling phase, the evaluation phase focuses on determining whether the models have properly achieved the business objectives specified in the earlier phases. This phase reminds students that data mining is not merely an academic exercise but a field designed to impact decision making, and it requires students to take off the hat of a technical expert and put on the business hat. One important process within the evaluation phase is to review the steps executed to construct the model to ensure that no important business issues were overlooked. Therefore, each team selects two students outside the team to review and critique the team's modeling process, and validate the logic behind the process.

This phase also impresses on students the importance of domain knowledge and business acumen in understanding the findings of analytics. During this phase, students frequently realize that the strongest patterns and relationships identified by the models are often obvious, less useful, or simply reflect business rules, and in many cases, the most technically elegant or sophisticated solutions yield little relevant insights to answer the business questions. In addition to self-assessment, student teams are asked to collaborate with domain experts to evaluate how well their predictive models achieve the business objectives. In the case of the ERIM project, each team conducts interviews with the instructor, who plays the role of a retail expert, to discuss the team's findings and explore the possible actionable decisions that retailers and manufacturers can make based on the findings.

For example, the classification models reveal interesting differences and similarities between households that are likely to purchase yogurt and those that are likely to purchase frozen dinners. Households that are likely to purchase yogurt tend to have higher income and education levels and consist of a married couple or are led by a female head of household; whereas households that are likely to purchase frozen dinners tend to have lower income and education levels and have at least one pet. Relatively young head(s) of households with large families are expected to purchase both yogurt and frozen dinners. Students are encouraged to develop compelling data stories that help depict the profiles of these households for the audience and provide actionable recommendations that would lead to marketing and advertising, store placement, and product design strategies. Such discussion often leads to teams backtracking to earlier phases to augment data preparation and modeling processes.

Putting on the business hat encourages students to look at the models from a different perspective sometimes. For example, while prediction models produce predicted values of the target variable, a marketing executive may decide to place more emphasis on the ranking of the predicted values rather than the values themselves. Similarly, in order to achieve our business objective, we are more likely to be interested in identifying households that would spend more on frozen dinners so that we can target these households for future marketing efforts rather than accurately predicting how much each household spends on frozen dinners. The performance measures such as RMSE often fail to provide us with this critical information. To understand the model's ability to correctly rank spending, performance charts such as the lift chart and the decile-wise lift chart can be more helpful. A critical evaluation of the analytics findings from the business perspective in this phase helps the teams refocus on the objectives of the project and create compelling data stories and recommendations for business decision makers.

Deployment

The final phase of the project involves a written report and presentation of the key findings. This phase stresses the importance of storytelling to communicate analytical findings to their intended audience effectively. Storytelling, or data storytelling, refers to crafting and delivering compelling data-driven stories to decision makers for the purpose of converting insights into actions, the final phase of the CRISP-DM framework.

Contrary to popular belief, storytelling is not the same as data visualization although presenting data through visually engaging figures and diagrams is a part of storytelling. Students are asked to focus on three key elements of storytelling: data, visualization and narrative, and how they complement one another to create a compelling story about the findings. Simply presenting the analytical process and findings from a technical perspective would have limited use to decision makers. To engage the audience, students must learn to focus on the context around the data that helps demonstrate the business value of the analysis, and use appropriate and engaging visualizations to help reveal the underlying patterns and relationships. Students are advised to present business insights gained from data analysis from a non-technical standpoint and craft the story around the data by focusing on answering the following three questions:

1. Why should the decision maker care about the findings?
2. How do these findings affect the business?
3. What actions do you recommend to the decision maker?

Storytelling gives the dry topic of data analysis an interesting spin and makes the content of the report and presentation more palatable to the audience who are often business decision makers with little training and/or interest in analytical methodologies. We find that storytelling is often intimidating at first to students in the business analytics course, especially to those with a technical mindset. However, it is a career-building skill that can be improved with practice and guidance from the instructor.

Critical reflection is an essential component of the experiential learning cycle (Kolb, 2015). It helps enhance students' understanding of the experiential activities in the context of the learning objectives of the course. Upon the completion of the project, the students are asked to reflect on their analytics work. A critical reflection framework such as the self-reflective model by Rolfe et al. (2001) can reinforce students' learning experience. Their reflective model is based on three simple steps: 'What?', 'So what?', and 'Now what?'. In the 'What?' step, students reflect upon important questions such as 'What happened in the project?', 'What was the role of each team member?', and 'What was the problem being solved?'. During the 'So what?' step, students may consider questions such as 'What other issues and opportunities arose from the project?', 'What conclusions did you draw from the project?', and 'What did you learn about the project and other team members?'. Finally, during the 'Now what?' step, students contemplate questions such as 'How will you apply what you learned from the project?', 'If you need to complete a similar project again, what would you do differently?' and 'What other skills might be beneficial to learn before you proceed to the next project?'.

ASSESSMENT OF PEDAGOGICAL EFFECTIVENESS

After completing an analytics course that provided students with a deep-dive of the CRISP-DM process, groups of graduate-level business analytics students then enrolled in an industry project course. In this course, students followed the six CRISP-DM phases to complete an analytic project addressing real-world business problems using large data sets. This sequence allowed us to assess the effectiveness of the CRISP-DM process as a pedagogical framework. Students filled out a course evaluation survey at the end of the 10-week quarter. The survey results were not released to the instructor until after the grades were submitted. Generally, students responded positively to the CRISP-DM pedagogical framework. Using the course evaluations from the same course that was offered in Spring 2017 where the CRISP-DM framework was not used as the pedagogical benchmark, we have seen improvements across multiple student satisfaction and learning measurements. In Winter 2019, approximately 89% of the students responded to the survey 'strongly agreed' or 'agreed' that the course was "educationally effective," whereas 83% students responded positively to this question in Spring 2017. In addition, 83% indicated that their interest in business analytics "has been stimulated" by the project versus 67% of the students in Spring 2017. Regarding the amount of workload, 89% responded that the workload was "appropriate in relation to other courses of equal credit" versus 83% of the students in Spring 2017. The students expressed that the CRISP-DM framework proved effective in helping them navigate through complex analytics issues and offered a structured approach to solving real-world, big data problems. While the anecdotal evidences from both students and the instructor regarding the effectiveness of the CRISP-DM pedagogical framework are extremely positive, more data will need to be collected in the future to validate our conclusion empirically.

At the end of the quarter, each student group presented its analysis results and key findings following the storytelling guidelines outlined in this teaching brief. The presentations focused on how the project delivered business value through a structured process of analytics with the intended audience being business executives rather than analytics professionals. Every student participated in the group presentation and was required to meet individually with the instructor to discuss his or her participation and contribution to the project. Students were graded on their mastery of the key knowledge points throughout the CRISP-DM phases and ability to communicate the findings effectively. For the Winter 2019 quarter, the oral presentations and individual meetings showed that approximately 82% of the students were able to articulate the business value of their projects and communicate the key findings to a non-technical audience effectively. Almost all of the students demonstrated a satisfactory level of understanding of the CRISP-DM framework.

CONCLUSION

Experiential learning opportunities that mimic real-world projects have been proven effective in teaching applied subjects such as business analytics. The project described in this teaching brief provides students with a holistic experience of converting data into insights and actionable business strategies. The infusion of the CRISP-DM framework throughout the project creates a structured approach to a creative problem-solving process. This extensive project is valued by both the instructor and students. The structure of the project and instructional experience gained by the instructor from this project can be readily applied to other large data sets and business problem contexts including consulting projects. Students who have undertaken this project gain a better understanding of the CRISP-DM framework, which is a widely adopted industry standard for analytics projects, and an integrated view of the knowledge points that permeate throughout the business analytics curriculum. The project provides an effective and engaging experiential learning activity that helps improve career-readiness for business students.

REFERENCES

1. Abbasi, A., Sarker, S., & Chiang, R.H. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, 17(3).
2. Anderson, J.S. & Williams, S.K. (2019). Turning Data into Better Decision Making: Asking Questions, Collecting and Analyzing Data in a Personal Analytics Project. *Decision Sciences Journal of Innovative Education*, 17(2), 126 – 145.
3. Asamoah, D.A., Sharda, R., Hassan Z.A., & Kalgotra, P. (2017). Preparing a Data Scientist: A Pedagogic Experience in Designing a Big Data Analytics Course. *Decision Sciences Journal of Innovative Education*, 15(2), 161–190.
4. Burch, G.F., Giambatista, R., Batchelor, J.H., Burch, J.J., Hoover, J.D., Heller, N.A. (2019). A Meta-Analysis of the Relationship Between Experiential Learning and Learning Outcomes. *Decision Sciences Journal of Innovative Education*, 17(3), 239 – 273.
5. Cardozo, R.N., W.K. Durfee, A. Ardichvili, C. Adams, A.G. Erdman, M. Hoey, P.A. Iaizzo, D.N. Mallick, A. Bar-Cohen, R. Beachy, & Johnson, A. (2002). Perspective: Experiential Education in New Product Design and Business Development, *Journal of Product Innovation Management*, 19(1), 4 - 17.
6. Dykes, B. (2016). Data Storytelling: The Essential Data Science Skill Everyone Needs. Forbes, March 31, 2016, available at https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/
7. Heim, G.R., Tease, J., Rowan, J., & Comerford, K. (2005). Experiential Learning in a Management Information Systems Course: Simulating IT Consulting and CRM System Procurement. *Communications of the Association for Information Systems*, 15, 428 – 463.
8. Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The Age of Analytics: Competing in a Data-driven World. McKinsey Global Institute, accessed June 13, 2019, available at https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

9. Institute for Operations Research and the Management Sciences (2019). Best Definition of Analytics, accessed May 3, 2019, available at https://www.informs.org/About-INFORMS/News-Room/O.R.-and-Analytics-in-the-News/Best-definition-of-analytics

10. Kolb, D. (2015). Experiential Learning: Experience as the Source of Learning and Development. New Jersey: Pearson Education, Inc.

11. Northwestern University (2019). Course Descriptions and Schedule, accessed May 8, 2019, available at https://sps.northwestern.edu/masters/data-science/program-courses.php?course_id=4790

12. Rolfe, G., Freshwater, D., & Jasper, M. (2001). Critical Reflection in Nursing and the Helping Professions: A User's Guide. Basingstoke: Palgrave Macmillan.

13. Rudin, C. (2012). Teaching 'Prediction: Machine Learning and Statistics'. *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

14. Silvester, K.J., J.F. Durgee, C.M. McDermott, & Veryzer, R.W. (2002). Perspective: Integrated Market-Immersion Approach to Teaching New Product Development in Technologically- Oriented Teams, *Journal of Product Innovation Management*, 19(1), 18-31.

15. University of Chicago (2019). Curricular Updates Spring 2018, accessed May 3, 2019, available at https://grahamschool.uchicago.edu/news/curricular-updates-spring-2018

16. Watson, H.J. (2013). Business Case for Analytics. *Biz Ed*, May/June, 49 – 54.

17. Wilder, C.R. & Ozgur, C.O. (2015). Business Analytics Curriculum for Undergraduate Majors. *INFORMS Transactions on Education*, 15(2), 180 – 187.

18. Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, Manchester, United Kingdom, 29 – 39.

APPENDIX A: SAMPLE BUSINESS ANALYTICS PROJECT ASSIGNMENT

This project is based on the data set called ERIMdata.xlsx that includes about 3,000 households in two midwestern cities in the United States. The data contain demographic information such as household incomes, number of household members, education levels of the heads of households as well as information on the purchases of a number of retail products such as frozen dinners and yogurt. The data were collected between 1985 and 1988 by a marketing research firm, AC Nielsen.

Your assignment is to first propose a business analytics plan based on the CRISP-DM framework and identify and complete the appropriate tasks for each of the six CRISP-DM phases. The project deliverables included in a final written report and an oral presentation should follow the outline shown below.

Business understanding: Describe the business opportunities that the data presents and formulate relevant business questions.

Data understanding: Explore the data set with descriptive analytics tools and provide relevant information. Examine the possibility of supervised and unsupervised analysis techniques and identify possible variables for further analysis. Keep in mind the business opportunities and questions formulated in the first phase. The following criteria may also be considered as a guide.

- Does a target variable(s) exist?
- Does the data set contain historical values of the target variable(s)?
- Does the data set have a sufficient number of observations to support data partitioning?

Data preparation: Determine and perform the necessary data wrangling and preparation tasks based on the decision made during the business and data understanding phases. Explain the rationale for these tasks and document the changes that you have made to the data set.

Modeling: Consider the strengths and weaknesses of different modeling techniques. Implement the appropriate techniques, explain the rationale for your selections, and present relevant analysis results and interpretation. For the supervised techniques, determine whether to use classification or prediction models and explain your decision. Use appropriate data partitioning and performance measures to evaluate the competing models implemented in the modeling phase. Identify the best model(s).

Evaluation: Refocus on the business objectives of the project. Review the steps executed to construct the model to ensure no key business issues were overlooked. Evaluate whether the models have properly achieved the business objectives outlined during the business understanding phase. Formulate actionable recommendations based on the findings.

Deployment: Communicate the findings and relevant business insights with a written report and oral presentation that incorporate appropriate statistical information and visuals. The main focus should be placed on providing actionable business recommendations for a managerial and non-technical audience.

APPENDIX B: DATA DICTIONARY

Table B1: Description of variables in ERIMData[1]

Variable	Description
HH_ID	The household's identification number
ResType	Types of residence: 1 for Apartment, 2 for Condo, 3 for Single Family, 4 for Multiple Family, 5 for Mobile, and 6 for Other.
ResStatus	Residence status: 1 for owned home, 2 for rented, and 3 for other.
HHInc	The average annual income of a household; there are 14 categories for this variable.
HHNbr	The number of members in the household.
MWrkHrs	The average hours worked each week by the male head of household.
MEdu	Education level of the male head of household: values less than 9 imply varying education levels prior to a college degree, 9 for graduated from college, 10 for attended graduate school, and 11 for post-graduate degree.
FWrkHrs	The average hours worked each week by the female head of household.
FEdu	Education level of the female head of household. See MEdu for detail.
FBirth	The birth year of the female head of household.
F_Rel	Relationship within the household: 1 for female head of household, 2 for male head of household, 3 for daughter, 4 for son, and 5 for other.
MBirth	The birth year of the male head of household.
M_Rel	Relationship within the household: 1 for female head of household, 2 for male head of household, 3 for daughter, 4 for son, and 5 for other.
Cable	Whether or not the household has cable; 1 if yes, 0 otherwise.
Cats	Whether or not the household has cats; 1 if yes, 0 otherwise.
Dogs	Whether or not the household has dogs; 1 if yes, 0 otherwise.
YogExp	A household's yogurt expenditures (in $)
DinExp	A household's frozen dinner expenditures (in $)

[1] Data are available upon request.

Table B2: Description of variables used for model development and analysis[2]

Variable	Description
HH_ID	The household's identification number
Cable	Whether or not the household has cable; 1 if yes, 0 otherwise.
HHInc	The average annual income of a household; there are 14 categories for this variable.
YogExp	A household's yogurt expenditures (in $)
DinExp	A household's frozen dinner expenditures (in $)
Yogurt	Based on YogExp, it assumes the value of 1 if a household purchases yogurt over the time period, 0 otherwise.
Dinner	Based on DinExp, it assumes a value of 1 if a household purchases frozen dinners over the time period, 0 otherwise.
Sglfam	Based on ResType, it assumes the value of 1 if a household resides in a single-family home, 0 otherwise.
Ownhome	Based on ResStatus, it assumes the value of 1 if a household owns a home, 0 otherwise.
Pets	Based on Cats and Dogs, it assumes the value of 1 if a household has a cat or a dog, 0 otherwise.
Married	Based on M_Rel and F_Rel, it assumes a value of 1 if a household has both a male and female head of household, 0 otherwise
EduBoth	Based on MEdu and FEdu, it assumes the value of 1 if both heads of households have at least a college degree, 0 otherwise
EduOne	Based on Married, MEdu, and FEdu, it assumes the value of 1 if a household led by a single person has at least a college degree, 0 otherwise.
WrkHrs	Based on Married, MWrkHrs, and FWrkHrs, it equals half of the hours worked by the two heads of the household or hours worked by the single head of the household.
HHMembers	Based on Married and HHNbr, it counts the additional members in a household excluding the head(s) of household.
Age	Based on MBirth and FBirth, it equals half of the total age of the two heads of the household or the age of the single head of household.
FHH	Based on Married and F_Rel, it assumes the value 1 if there is only a female head of household, 0 otherwise.

[2] Wrangled data are available upon request.

APPENDIX C: R CODE USED IN THE PROJECT

We first import the ERIMData into a data frame and label it myData. The following is a portion of the R code used for data wrangling and selected business analytic models.[3]

Data Wrangling

```
myData$Yogurt <- ifelse(myData$YogExp > 0, 1, 0)
myData$Dinner <- ifelse(myData$DinExp > 0, 1, 0)
myData$Sglfam <- ifelse(myData$ResType == 3, 1, 0)
myData$Ownhome <- ifelse(myData$ResStatus == 1, 1, 0)
myData$Pets <- ifelse((myData$Cats + myData$Dogs) > 0, 1, 0)
myData$Married <- ifelse(myData$M_Rel > 0 & myData$F_Rel > 0, 1, 0)
myData$EduBoth <- ifelse(myData$MEdu >= 9 & myData$FEdu >= 9, 1, 0)
myData$EduOne<-ifelse(myData$Married == 0, ifelse(myData$MEdu>=9 |
myData$FEdu>=9, 1,0),0)
myData$WrkHrs<-ifelse(myData$Married == 1, 0.5*(myData$MWrkHrs+myData$FWrkHrs),
(myData$MWrkHrs+myData$FWrkHrs))
myData$HHMembers<-ifelse(myData$Married == 1, myData$HHNbr-2,myData$HHNbr-1)
myData$Age<-ifelse(myData$Married == 1, (1985 -
0.5*(myData$MBirth+myData$FBirth)),(1985-(myData$MBirth+myData$FBirth)))
myData$FHH<-ifelse(myData$Married == 1, 0, myData$F_Rel)
myData = subset(myData, select = -c(ResType, ResStatus, HHNbr, MWrkHrs, MEdu,
FWrkHrs, FEdu, FBirth, F_Rel, MBirth, M_Rel, Cats, Dogs))
summary(myData$YogExp)
summary(myData$DinExp)
YogYes <- myData[myData$Yogurt==1, ]
YogNo <- myData[myData$Yogurt==0, ]
DinYes <- myData[myData$Dinner==1, ]
DinNo <- myData[myData$Dinner==0, ]
```

[3] The master code for a comprehensive list of analytical techniques, including the ones not detailed in the paper, is available upon request.

```
data.frame(colMeans(myData), colMeans(YogYes), colMeans(YogNo), colMeans(DinYes),
colMeans(DinNo))
data.frame(nrow(myData), nrow(YogYes), nrow(YogNo), nrow(DinYes), nrow(DinNo))
```

Classification Model (Logistic) for Frozen Dinners

Packages ‘caret’, ‘gains’, ‘pRoc’, must be installed and loaded before running the code.

```
myData$BinDin <- as.factor(myData$Dinner)
myDataD <- myData[, -c(1, 4, 5, 6, 7)]
set.seed(1)
myIndex <- createDataPartition(myDataD$BinDin, p=0.6, list=FALSE)
trainSet <- myDataD[myIndex,]
validationSet <- myDataD[-myIndex,]
Logit_Reg <- glm(BinDin ~ ., data = trainSet, family = "binomial")
Logit_Reg_Pred <- predict(Logit_Reg, newdata=validationSet, type = "response")
confusionMatrix(as.factor(ifelse(Logit_Reg_Pred>0.5,1,0)), validationSet$BinDin, positive =
'1')
confusionMatrix(as.factor(ifelse(Logit_Reg_Pred>0.33,1,0)), validationSet$BinDin, positive =
'1')
validation_BinDin <- as.numeric(as.character(validationSet$BinDin))
gain_table <- gains(validation_BinDin, Logit_Reg_Pred)
gain_table
plot(c(0, gain_table$cume.pct.of.total*sum(validation_BinDin)) ~
 c(0, gain_table$cume.obs), xlab="# cases", ylab="Cumulative", main="Lift Chart", type="l")
lines(c(0, sum(validation_BinDin)) ~ c(0, dim(validationSet)[1]), lty=2)
heights <- gain_table$mean.resp/mean(validation_BinDin)
dwlc<-barplot(gain_table$mean.resp/mean(validation_BinDin), names.arg = gain_table$depth,
ylim = c(0,2), xlab="Percentile", ylab="Mean Response", main="Decile-Wise Lift Chart")
r<- roc(validation_BinDin, Logit_Reg_Pred)
plot.roc(r)
auc(r)
```

Prediction Model (Regression Tree) for Frozen Dinners

Packages ‘rpart’, ‘rpart.plot’, and ‘gains’ must be installed and loaded before running the code.

```
myData$LnDin <- log(1+myData$DinExp)
myDataP<-myData[, -c(1, 4, 5, 6, 7, 18)]
set.seed(1)
myIndex <- createDataPartition(myDataP$LnDin, p=0.6, list=FALSE)
trainSet <- myDataP[myIndex,]
validationSet <- myDataP[-myIndex,]
default_tree <- rpart(LnDin ~ ., data=trainSet, method="anova")
summary(default_tree)
prp(default_tree, type=1, extra=1, under = TRUE, varlen = 10)
predicted_value_valid <-predict(default_tree, validationSet)
accuracy(predicted_value_valid, validationSet$LnDin)
gain_table <- gains(validationSet$LnDin, predicted_value_valid)
gain_table
plot(c(0, gain_table$cume.pct.of.total*sum(validationSet$LnDin)) ~
c(0, gain_table$cume.obs), xlab="# cases", ylab="Cumulative", main="Lift Chart", type="l")
lines(c(0, sum(validationSet$LnDin)) ~ c(0, dim(validationSet)[1]), lty=2)
heights <- gain_table$mean.resp/mean(validationSet$LnDin)
dwlc<-barplot(gain_table$mean.resp/mean(validationSet$LnDin), names.arg =
gain_table$depth, ylim = c(0,2.5), xlab="Percentile", ylab="Mean Response", main="Decile-
Wise Lift Chart")
```

Table 1: CRISP-DM phases and corresponding project learning objectives

CRISP-DM Phases	**Project Learning Objectives**
Business Understanding	Formulate business questions that lead to business strategies or actions.
Data Understanding	Describe the data in terms of the business context.
Data Preparation	Perform data wrangling to prepare the data for subsequent analyses.
Modeling	Develop predictive model(s) to inform decision-making.
Evaluation	Evaluate model performance and select the best predictive model(s).
Deployment	Communicate key findings through storytelling.

Table 2: Summary of yogurt and dinner expenditures

Expenditure	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**
Yogurt	0.00	1.20	10.33	40.60	35.84	3,258.40
Frozen Dinners	0.00	0.00	0.00	55.77	10.45	4,073.01

Table 3: Sample averages in all households and in subsets

Variable	**All**	**Yogurt**		**Frozen Dinners**	
		Purchase	**No Purchase**	**Purchase**	**No Purchase**
HH Income	6.0066	6.1805	5.2407	6.0668	5.9765
Cable	0.6579	0.6637	0.6322	0.6651	0.6543
Single Family Home	0.8736	0.8823	0.8356	0.9012	0.8598
Own Home	0.8498	0.8619	0.7966	0.8702	0.8396
Pets	0.5165	0.5321	0.4475	0.5823	0.4835
Married	0.7294	0.7430	0.6695	0.7883	0.6999
College Educated Both	0.1154	0.1301	0.0508	0.0988	0.1237
College Educated One	0.0618	0.0635	0.0542	0.0442	0.0706
Work Hours	27.4955	28.4386	23.3407	29.2324	26.6270
Other HH Members	1.0019	1.0843	0.6390	1.2775	0.8641
Age	48.2469	47.2564	52.6102	44.7662	49.9873
Female HH	0.2405	0.2336	0.2712	0.1910	0.2653
Number of HH	3189	2599	590	1063	2126

Table 4: Estimated logistic regression models for yogurt and dinner

Variable	**Yogurt**	**Frozen Dinners**
Intercept	0.3047 (0.447)	–0.0870 (0.814)
HH Income	0.0582 ** (0.007)	–0.0387 ** (0.025)
Cable	–0.0052 (0.959)	–0.0784 (0.349)
Single Family Home	–0.0614 (0.714)	0.1367 (0.365)
Own Home	0.3565 ** (0.022)	0.1359 (0.326)
Pets	0.0740 (0.456)	0.2186 ** (0.007)
Married	0.6901 ** (0.004)	0.4194 (0.109)
College Educated Both	0.6808 ** (0.001)	–0.4930 ** (0.000)
College Educated One	0.3977 * (0.073)	–0.2248 (0.250)
Work Hours	0.0014 (0.717)	–0.0045 (0.165)
Other HH Members	0.2455 ** (0.000)	0.1735 ** (0.000)
Age	–0.0095 ** (0.044)	–0.0214 ** (0.000)
Female HH	0.9031** (0.000)	0.1424 (0.585)

NOTES: Parameter estimates with the *p*-values in parentheses; * and ** represent significance at the 10% and 5% level, respectively.

Table 5: Performance measures of logistic models

Measure	**Yogurt**		**Frozen Dinners**	
	Cutoff = 0.5	**Cutoff = 0.82**	**Cutoff = 0.5**	**Cutoff = 0.33**
Accuracy	0.8141	0.5890	0.6541	0.5890
Sensitivity	0.9971	0.5833	0.1106	0.5671
Specificity	0.0085	0.6144	0.9259	0.6000

Figure 1: CRISP-DM life cycle (Wirth & Hipp, 2000)

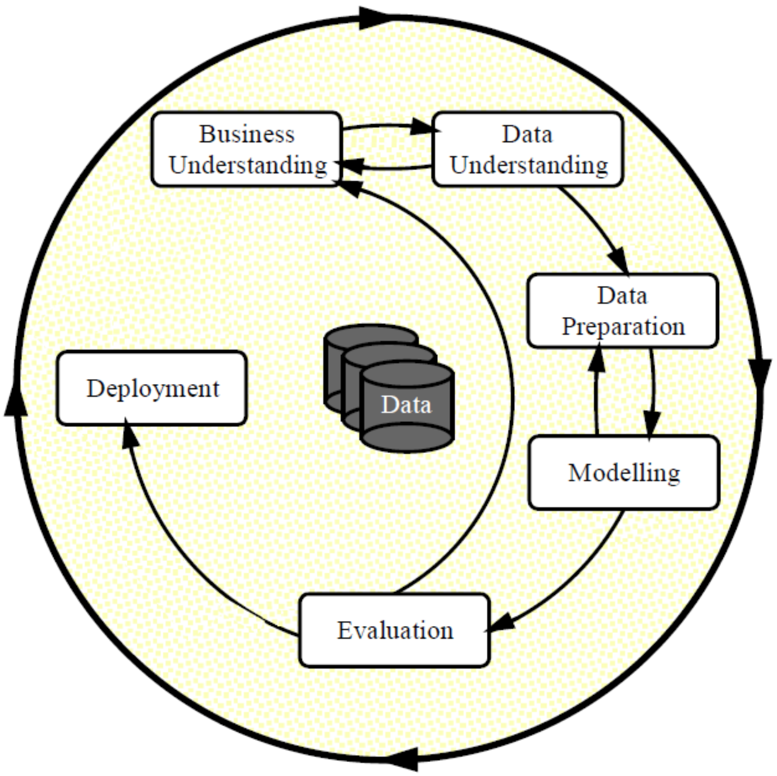
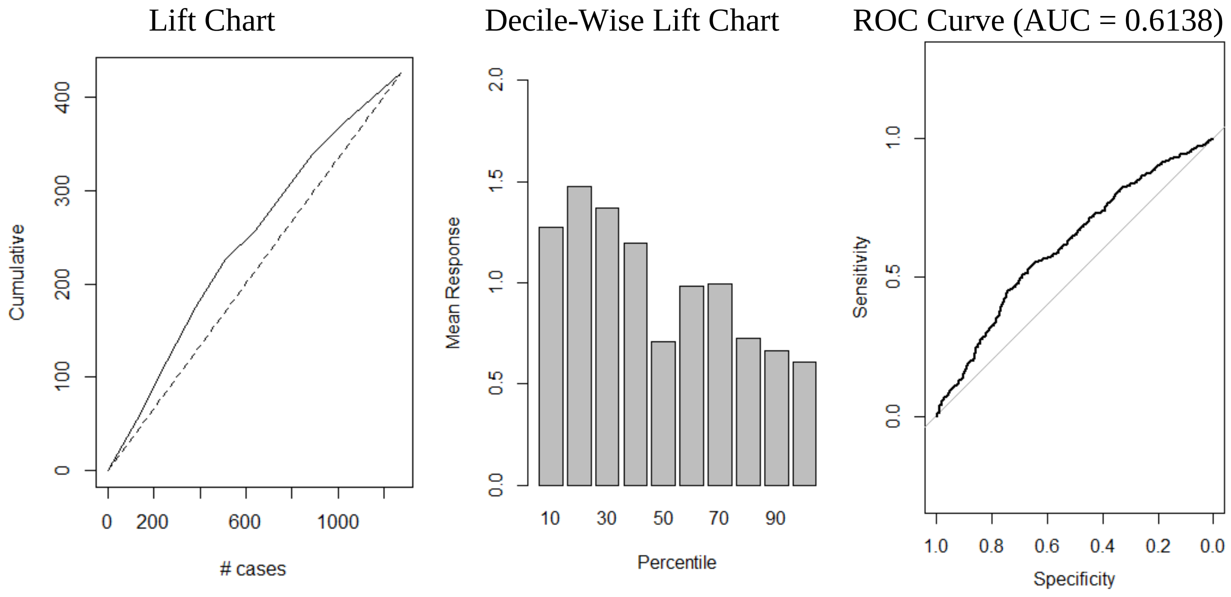


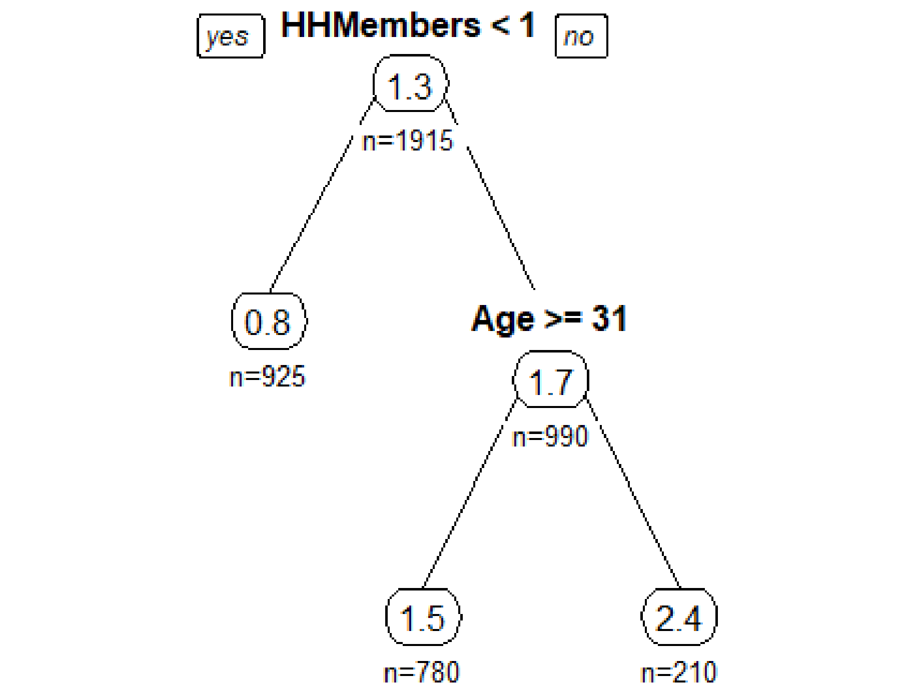
Figure 2: Performance charts for logistic classification for frozen dinners


Lift Chart

Decile-Wise Lift Chart

ROC Curve (AUC = 0.6138)

Figure 3: A regression tree model and its performance measures



ME	RMSE	MAE
-0.0048	1.9807	1.6155