KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics

Project Instructions

(in Teams of 3 to 5)

Prof. Espinosa – Last updated 12/20/2022

Table of Contents

Background

Business analytics is not just about analyzing data. Rather, it is about solving and then communicating solutions for business problems using credible data as your evidence. This requires teamwork, critical analysis, and a compelling upfront articulation of the specific business problem or analytics question you are solving as well as a clear and concise report of the findings and recommendations. Accordingly, the main goal of this project is to prepare for your practicum projects by giving you an opportunity to put into practice what you have learned in class.

The predictive analytics project will be done in teams of approximately four students. We expect that all team members will contribute equally and that everyone will take the opportunity to learn from each other.

CRISP-DM Overview

The **Cross Industry Standard Process for Data Mining** (**CRISP-DM**) is the most widely used networks for data mining and analytics projects. I provide an overview of this network as an informational piece. The CRISP-DM maps closely to the **INFORMS' Job Task Analysis** (**JTA**) framework (http://info.informs.org/analytics-body-of-knowledge; Amazon), which is also followed widely in management science projects. The deliverables for this project are largely based on a shorter version of the CRISP-DM framework, so this will provide you with a good reference frame. In essence, the CRISP-DM framework and the corresponding JTA activities and domains include the following activities:

- **Business Understanding** (CRISP-DM 1)
  - (*JTA Domain I*) Formulate the *business question* to be answered or problem to be solved. All business analytics projects must be driven by business needs or business value propositions. This requires clearly stating the business question or problem you hope to solve upfront.
  - (*JTA Domain II*) Translate the business question into the respective *analytics question*. Not all business questions or problems are amenable to analytics solutions. The project report must specify how or why analytics is the appropriate approach to address the business question or problem.
- **Data Understanding** (CRISP-DM 2)
  - Data acquisition and pre-processing can take as much as 80% of the analytics project effort.
  - (*JTA Domain III*) Acquire and identify relationships in the data. This step involves *acquiring the data* (e.g., ETL or Extract-Translate-Load) and then doing a substantial amount of *descriptive analytics*, including things like (as appropriate): descriptive statistics, correlation analysis, ANOVA, distribution curves, visual plots and other graphs, and other related analysis (e.g., cluster analysis). NOTE: Predictive analytic modeling should not start until you have developed a thorough understanding of the data. If fact, this phase may uncover issues and relationships in the data that you did not anticipate, thus leading to reformulation of the analytics question.
- **Data Preparation** (CRISP-DM 3)
  - (*JTA Domain III*) Harmonize, re-scale and clean data, as needed. Data sets often need to be split, merged, sub-sampled (for large data sets), and cleansed. This step involves all *data pre-processing* activities, such as: re-structuring the data (e.g., normalizing scales, centering, aggregating, etc.); addressing issues of missing data; and acquiring and merging other related data.
- **Modeling** (CRISP-DM 4)
  - Select the appropriate analysis *methodology* and tools, exploring various model *specifications*, and then building the respective models. In this course we use R as the primary analytical tool.
  - (*JTA Domain IV*) Methodology Selection. Most of the course is focused on *method selection* (e.g., OLS regression, Logistic regression, Ridge or LASSO, trees, etc.). Candidate models should be identified based on the analytics goals: interpretation, inference and/or prediction). For this project, students need to focus on models that are relatively

interpretable and then select the model that has better predictive accuracy, based on cross validation test error or deviance.

  - (*JTA Domain V*) Model Building. Another area of focus in this course is on *model specification,* e.g., linear, polynomial, interactions, or variable selection. The initial set of predictors to be used in the model must be driven by business domain knowledge. But then this set should be narrowed down or refined using statistical methods like cross-validation testing.

- **Evaluation** (CRISP-DM 5)
  - Note that this phase is not about evaluating the models. This happens in the Modeling phase above. **This phase is about evaluating the extent to which the analysis has answered the business and analytics questions framed in phase 1**. For this project, we will focus on the following:
  - Interpretation of Results: the final project reports must provide a focused *interpretation of results*, in terms of effects observed, fit statistics, and predictive power of the final model.
  - An important part of this interpretation is providing a well-documented answer to the business and analytics question - how will these findings be applied or drive strategy?
  - It is also important that you tell a compelling story in your report. *Storytelling* is one of the most important skills in business analytics. Remember, this is not a statistics class, but a business class. You must tell a compelling story for your audience. The story must be backed up by your findings.

- **Deployment** (CRISP-DM 6)
  - (*JTA Domain VI*) For this project, deployment will focus on turning in your *written report*, with the necessary interpretation and stories articulated in step 5 above.

Important note: not all projects lead to amazing findings!! A model that shows no effects can offer very interesting insights. It all depends on how you rationalize the lack of effects from a business point of view. Along the same lines, this project is not so much about what you analyzed and found, but about how effectively you described to your readers the motivation for your study, your method evaluation and selection process and what the implications of your findings are from a business perspective.

Data

Any dataset not used in class for lectures, exercises, homework or projects by students in prior semesters can be used for this project. Students are expected to identify an interesting external data set to work with. In the past, many students have used Kaggle data sets used in competitions, but there are many sources of public data. Proprietary data sets can only be used with permission of the owner of the data set. It is fine to use data from your practicums, if you have it, and use this project as an opportunity to work with your client's data. Unless the data is proprietary, teams must submit the actual datasets with their final projects so that the professor can replicate some of your work when grading. If you use a dataset that you used previously in another course, you need to provide the name of the course and instructor. You also need to briefly describe what you did in the earlier project and your analysis for the current project must be substantially different than what you did in that course.

Requirements

- **Teams** of **3** or **4** students **must** evaluate **3** different **modeling methods** (e.g., OLS, Ridge, Logistic, LDA, trees, etc.) with **2** different **model specifications** for **each** method (e.g., different predictor subsets; polynomial, log or other transformations; interactions, etc.).
- **Teams** of **5** need permission from me and must evaluate either **3** different **modeling methods** with **3** different **model specifications** for each method; or **4 modeling methods** with **2 model specifications**. There are three critical points to remember:

POINT #1: The 2 model specifications selected above should be used in **each** of the 3 modeling methods above. The best approach is to fit the first model using OLS for a quantitative prediction, or Logistic regression for a classification prediction, using both model specifications. Then, depending on your results and assumption testing, fit the same 2 specifications using two other models.

POINT #2: All team members must contribute their fair share of the analysis. The expectation is that each member will take the lead on one particular modeling method or transformation. Each team will provide feedback on their teammates during the semester to evaluate how each member contributed to the project. I will ask students to report on their individual contribution to the team's project.

POINT #3: While you will be evaluating and testing 6 different models (3 model methods x 2 specifications), you should **only report on the final model methods and specification selected**. You only need to discuss your model selection process, including any fit statistics and cross-validation test results for all 6 models and discuss how you arrived to your final selection. Once you select a particular modeling method and specification, you need to fit your final model with the **full dataset.** If you wish to include output from alternative models and specifications, you can do that in an appendix.

POINT #4: You don't need to include all your R code and output. There is simply no time to review all your R code and output. But you can provide some of the key output in appendices, if you wish.

POINT #5 – IMPORTANT: All your deliverables must contain subsections matching the items requested below. Please don't submit a monolithic report with no section titles. It only makes it more difficult to evaluate your project, so help me give you the credit you deserve for your good work, rather than guessing.

Modeling Method, Model Specification and Procedures

Analytics Question	Model Types	Modeling Methods (Evaluate 3)	Model Specification (Evaluate 2 for each model method)	Testing, Procedures and Comments
Quantitative	Parametric	• Regression: OLS, Ridge, LASSO, PCR, PLS	• Predictors (feature) selected based on business knowledge • Predictors (features) selected by statistical analysis (e.g., variable selection, ANOVA, stepwise, etc.) • Variable transformations (e.g., polynomials, log models, lagged or time-series models, piecewise models, splines, ranks, Box-Cox transformations, etc.) • Categorical to binary variable transformations are not different specifications but required transformations. • Centering and standardizing are not model specifications, but different interpretations of results.	• WLS is not really a different method, but a correction for OLS' heteroskedasticity. • GLM is not really a different method, but an estimation method applicable to several methods. • Step variable selection (Forward, Backward and Stepwise) are not model methods, but evaluation procedures for variable selection. • Bootstrap, RSCV, LOOCV and KFCV are not model methods, but cross-validation tests applicable to any method. • You can use R^2, Adjusted R^2, model p-value, coefficient p-values, Cp, 2LL, AIC, BIC, etc. to analyze the statistical fit. But all candidate models and specifications must be compared with a cross-validation testing method. • Changing tuning parameters don't yield different models or specifications but help select the optimal tuning of models (e.g., tree size, Ridge shrinkage, PCR components).
	Non-Parametric	• Regression trees, bootstrap aggregation trees, random forest trees, boosted trees, neural networks, k-nearest neighbors		
Classification	Parametric	• Binomial and multinomial Logistic regression • LDA and QDA		
	Non-Parametric	• Classification trees, bootstrap aggregation trees, random forest trees, boosted trees, support vector machines, neural networks		

Note: only models and specifications covered in this course are allowed for the course project. Any team contemplating other methods not covered in the course must obtain permission from the instructor.

Project Deliverables

Project Template

This project must be completed using the project template provided in Canvas. Use the same template and project document for all deliverables. The idea is to complete your project incrementally in a single document. This template contains all the required sections for the project. And departures from the template must be discussed with and approved by me.

This project has **5 deliverables**:

Deliverable 1 (10 pts): Project Proposal (1 page, single-spaced or 2 pages double-spaced)

A project proposal is due around the mid-semester point, per the class schedule. The goal in this deliverable is to get you started on your project early and provide the direction you are planning to take in your project. It is also an opportunity to get feedback on your project ideas.

Previously used data: If you use a dataset that you used previously in another course, you need get approval from me. You also need to provide the name of the course where you used the data and the instructor's name. For the approval, you also need to briefly describe what you did in the earlier project in your proposal and how your analysis for the current project will differ from the past project.

The proposal should contain the following sections:

(1) **The business question and case** – you need to articulate both.

   a) The **business question** is the one you will be pursuing in your project, articulated in a way that will resonate with a managerial audience and it tends to be a larger and more general question (e.g., how can the spread of the current epidemic be controlled?).

   b) The **business case** should follow the business question. It should contain a brief rationale about the importance of this question/problem from a business perspective. Why is it important for your managerial audience? What is the value proposition of your project?

(2) **The analytics question** –not all business questions can be answered with analytics. The business question for this project must be answerable through analytics. The analytics question is more technical in nature and should contain a **plain English** articulation of:

   a) The **outcome** variable of interest for the study, with a clear indication of type of variable it is – i.e., **quantitative** (continuous or count) or **classification** (binary or categorical); and

   b) The **key predictors** of interest in your model. You don't need to discuss all possible Predictors, but just the **focal predictors** of interest at the time of the proposal.

   c) A brief articulation of how answering this analytics question helps answer the business question above.

(3) **One or more possible datasets** identified for the project. The more specific the datasets you are contemplating the better. You need to specify the possible sources for your dataset (e.g., Kaggle, Google Datasets, etc.)

Note: if your team wishes to change the proposal after it has been submitted, it is OK, but you will need to prepare and submit a new proposal and any other deliverable you may have already submitted. You can include the revised proposal in the Deliverable 2 document below.

Deliverable 2 (10 pts): Preliminary Data Analysis Report (2 pages of text **MAX**, single-spaced, or 4 pages double-spaced, plus appendices with R output as needed)

This deliverable is intended to get you started early on your project model method and specification exploration. It is also meant to get you familiarized with the project data. You should think of this deliverable as an **early draft** of your final report. It is also one last opportunity to get feedback on the direction of your project.

Because all model explorations begin with either an ***OLS*** regression (for **quantitative** predictions) or a ***Logistic*** regression (for **classification** predictions), this preliminary data analysis report will include the following:

(1) **IMPORTANT:** your main text should contain only narratives. Place all statistical output and plots in appendices. All appendices must be appropriately referenced in the main text. Only include appendices and R output if they support your main text narrative. Unrelated appendices and output are confusing and unnecessary.

(2) **Revise and refine your project proposal as needed.** More specifically, refine your business case, business question and analytics question, as needed. Your deliverable 2 report **must include** these revised items. Everything in your proposal (i.e., business case, business question, analytics question, etc.) should also be part of your deliverable 2 report. That is, build on your proposal. Don't write a new document from scratch.

(3) **Brief description of your dataset.** In Deliverable 1 you discussed possible datasets to use. For this deliverable, you must have settled on the specific dataset you will use in your project. You don't need to provide a full description of the dataset yet, but you need to provide enough information for your professor to understand what you are analyzing. No need to provide extensive descriptions, just the **data source**, the **main variables** of interest to your study contained in the data set, the total number of observations and, for each variable please describe its respective variable type, unit of measurement, and a brief description.

(4) **Descriptive analytics.** You must provide a brief discussion of the respective descriptive statistics, correlation analysis, ANOVA and/or any plots you may have rendered to understand the data and how variables relate to each other. You should not only discuss the types of analysis you did, but why you did them and what you learn from them to inform your upcoming predictive modeling.

(5) Define an initial **set of predictors** for your model. These predictors must be variables in your dataset and must be selected using business domain rationale. The initial set of predictors should NOT be selected statistically, but you must articulate your rationale for why you chose your initial set of predictors.

(6) If your analytics question is **quantitative**, run an **OLS regression**. Or, if your analytics question is a **classification**, run a **Logistic regression**. In either case you must include the predictors

identified above. Later in the project you will refine this initial set of predictors through variable selection, best subsets, or other methods.

(7) **Inspect** residual and other regression **plots**, as appropriate. If your model is quantitative, conduct the necessary assumption **tests** to evaluate adherence to the OLS regression **assumptions,** (e.g., multicollinearity, serial correlation if there is time data, heteroscedasticity, linearity). If your model is for classification, you only need to test for multicollinearity and serial correlation if there is time data.

(8) Provide a brief statement of your conclusions. You must discussed, what types of possible models you will evaluate next and what did you learn in general.

Deliverable 3 (5 pts, individually graded): Meet with Professor

The full team needs to meet with me shortly after the Deliverable 2 submission. This meeting is required. Team members will be graded individually based on their understanding of the project issues, general contributions and participation in the Deliverable 3 meeting. In addition to these individual points, the final project grade may be adjusted upwards or downwards based on team evaluations and my evaluation of your contributions towards the project. ALL team members must attend this meeting, so please schedule your meeting with me ahead of time. A team member that does not attend this meeting will not earn the assigned points above. The purpose of this meeting is for me to learn how your team is working out, what is the general contribution of each team member, discuss any methodological issues or challenges you may have encountered, and for me to give you feedback on your Deliverable 2 and your project as a whole.

Deliverable 4 (65 pts): Final Report (4 to 5 pages of text, single-spaced, plus appendices with R output as needed)

IMPORTANT: as it should be clear by now, one important learning objective in the MS Analytics program is being able to interpret analytics results and articulate them clearly to a business audience. The market calls this **"storytelling"** and it boils down to writing concisely, to the point and clearly what your results mean for your client. "Storytelling" means that you are persuasively communicating information to address a problem the client faces and could be "a story" about agile solutions, concern for brand reputation and efficiency, gaining ground over the competition, or otherwise improving the business practices of a client to their advantage. In today's business environment, your role is critical, because all business "stories" must be backed up by evidence, such as your interpretations of statistical output. Avoid grandiose statements and fluff. Get to the point right away because the space is limited and business people like succinct but informational writing. Without a story, simply presenting data does not drive effective business decision-making.

The final project report will be submitted as an analytics report prepared in MS Word or knitted with R Markdown as a Word or PDF document. Most of these sections should be an extension of your Proposal and Preliminary Data Analysis Report above. The final project report will contain the following sections:

(1) The **business question** and **business case** (5 pts.), as described above. Please update this section as needed.

(2) The **analytics question** (5 pts.) being addressed, as described above. Please update this section as needed.

(3) A description of the **dataset** (5 pts.) utilized for the analysis (if the data set is not available in an R package or public web site, the data source must be indicated with a specific link). Your data description should be sufficient for your reading audience to understand your data set, variables and the interpretations you provide in your report, including **variable types** and **units of measurement** (especially for quantitative variables) and **number of observations** in the data set. The data description should be accompanied by any necessary descriptive analytics artifacts necessary for your predictive modeling, e.g., descriptive statistics, correlation matrix, correlation plots, or other plots. There is no need to list all variables in the data set and this is NOT a place to discuss your variable selection. It is recommended to provide a narrative describing the key study variables and refer to a table in the appendix with all the variables. Please note that your dataset may contain hundreds of variables, but you only need to describe variables that are relevant to your analysis.

(4) **Descriptive Analytics** (10 pts.): Brief analysis of the study variables, from both business and statistical perspectives. Your descriptive analytics must address the following:

a. **Variables:** Provide a list of all variables (i.e., outcome and predictors) you employed in your study. For each variable: (i) if it is the outcome variable, no need to describe it further; (ii) if it is a main predictor, provide a brief rationale about why you think these predictors influence the outcome; (iii) for other predictors, just provide a one line reason for why you included them in the model (i.e., typically, these would be your control variables).

b. **Descriptive Analysis:** Analyze your main variables (i.e., outcome and main predictors). It is customary to describe the means, variances, correlations and visual graphs and plots for the key variables in the model, although other descriptive statistics could be useful.

c. **Observations:** Briefly discuss any important aspects uncovered by your descriptive analytics of the data (i.e., visual plots, descriptive statistics, correlations, etc.). The purpose of descriptive analytics is to develop familiarity with and a better understanding of the data and possible patterns. What did you learn?

d. **Assumption Tests:** Also, provide a brief report of the OLS or Logistic model assumptions you tested and the respective results. You don't need to test for everything. Be selective and provide a brief rationale for why you decided to conduct a particular test. Note that you will need to fit an OLS or Logistic model to do this, but at this time you are not evaluating or interpreting these models, just testing assumptions.

e. **Data Pre-Processing:** Finally, provide a brief discussion of any **pre-processing,** e.g., grouping, combining variables, etc., and **transformations** done with the data, e.g., normality, logs, standardization, non-linear, or other, you employed or plan to employ for some of the main variables, if any, along with a brief **rationale** for the appropriateness of this transformation (e.g., normality, non-linearity, non-continuous, etc.). Again, you will be selecting your model specifications later, but you want to do some descriptive analytics early to spot any issues with the data that may require transformations.

Please include all the necessary plots, descriptive statistics, correlation matrices, etc. in an **Appendix**. DO NOT include all your R output, but only the output that supports your narrative. Also, do not include R output in the main text, but in appendices. And most definitely, DO NOT include your R scripts.

(5) A **discussion** of the (a) **analytics methods** and (b) **model specifications** you evaluated and selected (10 pts.). All methods used must be appropriate and relevant to the problem and you need to provide a justification for the selected methods based on:

a. **Model Candidates:** besides **OLS** or **Logistic regression**, provide a brief discussion of your model candidates, based on your descriptive analytics and assumption tests you conducted above.

b. **Model Specification Candidates:** based on your variable selection and data pre-processing results above, provide a brief description and discussion of the model specifications you are considering.

c. **Cross-Validation Testing and Final Model Selection:** provide a brief description of the cross-validation test results from all the candidate models and specifications you considered, and indicate your final model selection and the criteria you based your decision on.

(6) Analysis and presentation of **results** (10 pts.). Your analysis and results need to contain some narrative to allow your audience to understand what you did. A simple output, diagram and data dump with no explanation will receive very little credit. Every procedure, output and diagram need to be briefly but appropriately introduced before and briefly commented on its meaning after. Don’t leave it up to the reader to interpret what you did. Also, vague and general discussions of results will receive little credit. Your narrative of results should be factual and specific, so it needs to be backed up by fit statistics, coefficient values and significance, etc. Most importantly, your analysis must address the **analytics question** you articulated above. That is, you start your report by formulating a question you intend to answer, so therefore you must end by answering that question.

(7) A short section with **final thoughts**, **conclusions** and **lessons learned** (10 pts.). Business analytics is about gaining insights from business data for decision making. This is the section for you to articulate what insights you gained from your analysis. These **conclusions** must contain a discussion of:

a. The main **conclusions** of your **analysis**. These conclusions must answer/solve your **business question/problem** stated in 1 above. Please be brief but concise and discuss the main insights obtained from your analysis

b. A brief statement of the main issues and challenges you faced in this project and what you learned from it, including things like: data issues, methodological challenges, do's and don'ts, what you learned from this experience. You don't need to address all of this. But please be thoughtful and make it interesting.

(8) **Writing** Quality, Formatting and Presentation (10 pts.). We have heard over and over from recruiters, practicum managers and our ITEC advisory board that **storytelling is everything**. Managers are less interested in p-values and other fit statistics, and more interested in the meaning and insights you gained from your analysis, and how these influence business decisions and outcomes. Analytics projects, no matter how good they are, are not useful unless the analytics report is well written and clearly articulated. Statistical output without sound commentary about the results and implications for business are of little use to decision makers. Consequently, the insights from your story, along with the businesslike document formatting, presentation, writing clarity of the report, free of grammatical errors and typos will be heavily considered as a critical part of your grade. Just as importantly, the entire report needs to flow, persuade, and be understandable to your audience.

Deliverable 5 (10 pts): Brief Presentation to the Class (5 to 6 slides of content)

Each team will have approx. 10 minutes (including Q&A) to present the project to the class your: business question/problem; model selection; and conclusions. All presentations must follow this format (approximately one slide per each bullet):

- Title slide with project name and team members names
- Business problem, business case and analytics question addressed in the study.
- Brief description of the dataset (describe any relevant aspects of descriptive statistics, correlations, visual plot inspections, and pre-processing or transformations, as appropriate)
- Brief explanation of your model selection process and alternatives, along with the respective model specifications.
- Discussion of the final model and the most relevant results. No need to discuss all results (there is not enough time), just important ones.
- Final conclusions about implications of your findings. Are there recommendations for client action that you can make based on the data?
- Brief articulation of the challenges you encountered in your project. Discussing the lessons learned and actionable steps informed by your data insights is a useful way to link your work back to the client, the story, and the strategy. As we stated earlier, effectively communicating data insights to solve a business case is why you as data analysts are in such demand in the professional world.