KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics

Predictive Analytics Project

Prof. Espinosa – Last updated 12/30/2018

Background

The main goal of this project is to help you 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 maximum 4 students. It is expected that all team members will contribute equally and that everyone will take the opportunity to learn from each other. Business analytics is not just about analyzing data. It requires a compelling articulation upfront of the specific business problem or analytics question being addressed; and a clear and concise report of the findings and conclusion.

Students will formulate a business problem (or question) to address (or answer) through predictive analytics. The goal is for students to: identify an appropriate dataset to analyze the problem/question; evaluate a number of suitable models and model specifications; select the most appropriate method model specification; and apply them to answer the business question. Students will identify potential use of predictive analytics, formulate the problem, identify the right sources of data, analyze data, and prescribe actions to improve not only the process of decision making but also the outcome of decisions.

Important: not all projects lead to amazing findings. A model that shows no effects can be an interesting finding. 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 from a business perspective.

Data

Any dataset not used in class for lectures, exercises or homework 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 OK 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.

Requirements

All projects **must** evaluate **3** different **modeling methods** (e.g., OLS, Ridge, Logistic, LDA, trees, etc.) with **2** different **model specifications** for **each**, (e.g., different predictor subsets; polynomial, log or other transformations; interactions, etc.).

IMPORTANT: all team members must contribute their fair share of the analysis. I expect each member to take the lead on one particular modeling method or transformations. I will be surveying the team during the semester to evaluate how each member contributed to the project.

IMPORTANT: 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, but you must close the loop and re-fit your final model with the **full dataset.** There is no need to report on all alternative models. You only need to discuss your model selection process, including any fit statistics and cross-validation test results that guided your final selection. However, if you wish to include output from alternative models and specifications, you can do that in an appendix.

Project Deliverables

This project has **4 deliverables**:

Deliverable 1 (5 pts): Project Proposal (1 page, single-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 me with an idea of the direction you are planning to take in your project. It is also an opportunity for me to give you feedback on your project ideas. The proposal should contain the following sections:

(1) The analytics question/problem being addressed
(2) A brief rationale about the importance of this question/problem from a business perspective
(3) One or more possible data sets identified for the project; and
(4) A discussion of a few tentative predictive modeling methods you may employ (which you can change later if needed; also, no need to provide model specifications this early).

Note: Often teams change projects after submitting the proposal. This is perfectly fine, but it requires that you prepare a new proposal for the re-formulated project.

Deliverable 2 (10 pts): Preliminary Data Analysis Report (2 pages of text, single-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. While it is not required to meet with me, all teams are encouraged to schedule meetings for me throughout the semester to discuss your projects.

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) Analytics question/problem formulation, which will be the entire basis for your project. You need to clearly specify if your question is about predicting a quantitative or classification outcome, or perhaps both.

(2) Articulation of your analytic goals: interpretation, inference and/or prediction

(3) Brief description of your dataset. 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 and the main variables included in your analysis (please explain their respective variable types and metrics).

(4) Brief discussion of the respective descriptive statistics and correlation analysis. The text in this section should be limited to a brief analysis of the most salient aspects of this analysis, but do not include the actual de descriptive statistics and correlation matrix in the main text. But you must include them in an appendix.

(5) If your analytics question is **quantitative**, run an **OLS regression**. If your analytics question is a **classification**, run a **Logistic regression**. In either case you must include the most appropriate variables, selected with either **best subset** or **stepwise methods**.

(6) **Inspect** residual and other regression **plots**, as appropriate, and conduct the necessary **tests** to evaluate adherence to the OLS or Logit regression **assumptions** (e.g., multicollinearity, serial correlation if there is time data, heteroscedasticity, linearity, etc.).

(7) Provide a brief statement of your conclusion.

Deliverable 3 (75 pts): Final Report (4 to 5 pages of text, single-spaced, plus appendices with R output as needed)

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

(1) (10 pts.) The **analytics question/problem** being addressed. In your framing of the analytics question, please state clearly articulate:

   a) The specific predictive analytic question you are attempting to answer, or problem you are trying to solve? Since this course is about predictive analytics, it is important that you clearly specify the response variable you are predicting.

   b) The type of problem/question you are addressing – i.e., quantitative, classification or both

   c) Your project's analytics goal(s) – i.e., inference, interpretation and/or prediction.

(2) (5 pts.) A brief but compelling **business case** articulating the **rationale** about the importance of this question/problem from a business perspective. Why is the problem you are analyzing important?

(3) (10 pts.) A description of the **dataset** utilized for the analysis (if the data set is not available in an R package or public web site, the data set must be attached). 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. 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, other plots, etc.).

(4) (10 pts.) **Descriptive Analytics**: Brief analysis of the study variables, from both, business and statistical perspectives.

a) First, clearly identify and describe your outcome variable(s).

b) Then specify and briefly describe your main predictors. You don't need to discuss all predictors in this section, just the ones that are most central to your analytics question and business problem. You will be selecting the final predictors later, but before you do that, it is important to have a business rationale for including them.

c) Briefly discuss any important aspects uncovered by your descriptive analytics of the data (i.e., visual plots, descriptive statistics, correlations, etc.)

d) 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, etc.) you employed for some of the variables, if any, along with the **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 R output in the main text.

(5) (10 pts.) A **discussion** of the (a) **analytics methods** and (b) **model specifications** you evaluated and selected. 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) Conformance with or departure from OLS and/or Logistic OLS assumptions, based on visual inspections and OLS assumption tests.

(b) Predictive accuracy based on **cross-validation test** statistics. Similarly, the particular model specifications utilized must have a rationale. For example, if you chose a quadratic regression specification, you must have some rationale for the respective non-linear relationship. All projects must be analyzed with a variety of appropriate model with different model specification. Please consult with me if in doubt, but these are the minimum requirements

(6) (10 pts.) Analysis and presentation of **results**. Your analysis and results need to contain some narrative to allow your audience to understand what you did. A simple output and diagram dump with no explanation will receive very little credit. Every procedure, output and diagram needs 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 backed up by fit statistics, coefficient values and significance, etc.

(7) (10 pts.) A short section with **final thoughts**, **conclusions** and **lessons learned**. 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:

- The main **conclusions** of your **analysis**. These conclusions must answer/solve your a**nalytics question/problem** stated in 1 above. Please be brief but concise and discuss the main insights you obtained from your analysis
- 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) (10 pts.) **Writing** Quality, Formatting and Presentation. Analytics projects, no matter how good they are, are not useful unless the analytics report is well written and clearly articulated. Nobody wants to see a bunch of statistical output without sound commentary about the results and their implications for business. Consequently, heavy weight will be placed on the attractiveness, presentation, writing clarity of the report, free of grammatical errors and typos. More importantly, the entire report needs to flow and be understandable to your audience.

Deliverable 4 (10 pts): Brief Presentation to the Class (5 to 6 slides of content)

Each team will have 10 to 12 minutes or so to share with 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 or analytics question addressed in the study with a short statement of the business case
- 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 most relevant results. No need to discuss all results, just important ones.
- Final conclusions about implications of your findings
- Brief articulation of the challenges you encountered in your project.