BUS256 Marketing Analytics
Fall 2018

Final Project: Investigating Data & Producing Managerially Relevant Decisions

Final Report

Introduction: The final report for the team project extends the preliminary analysis from the first two deliverables, and involves conducting in-depth predictive analyses on the chosen dataset. The objective is that your team understands the broad managerial goal for the project you have chosen and links it with the results of the statistical analyses.

Datasets: Your team has chosen one of these four datasets –

- Synthetic Financial Data from PaySim for Fraud Detection (https://www.kaggle.com/ntnu-testimon/paysim1/home)
- Analyzing Black Friday Purchases for customers and product categories (https://www.kaggle.com/mehdidag/black-friday/home)
- Analyzing the viability of Kickstarter projects (https://www.kaggle.com/kemical/kickstarter-projects/home)
- 2015 Flight Delays and Cancellations – which airline to fly on?(https://www.kaggle.com/usdot/flight-delays/home)

For detailed description of these datasets, please read the Team Project description document and the associated Kaggle webpages.

Assignment: The data assignment and tasks depend on the project chosen. Your report should build on your previous submission for Interim Deliverable II. Certain questions or sub-parts may seem repetitive. These are an opportunity for you to incorporate instructor comments for your previous submission and improve on your report. A brief description of the assignment for each dataset is below.

In terms of report structure, your report should have **six** components –

1. **Introduction –** This should briefly talk about the data problem, why is it interesting to look at this problem (i.e. managerial objective), and the broad goals of your project.
2. **Data Description –** This should “introduce” the dataset to the reader. It should cover the following points –
   a. Describe the “conceptual” measure types of the different variables in your data.
   b. Data Cleaning – Mention all the steps you took to clean the data. This could include changing the computer data-types of the variables (type coercion), dealing with missing data, filtering out observations, selecting variables, etc. *All*

the datasets that you use later for further analysis should be generated in this section.

3. **Summary statistics and Data Visualizations** – This would describe the basic patterns in your CLEANED dataset. Specifically, apart from showing the code and resulting output, you should explain the following -
    a. What question are you trying to answer with each summary table or visualization?
    b. What conclusion do you draw about the answer to your question?

    You can add details about the following in your description –

    a. Why did you select the variables (and summaries) that you chose?
    b. Why did you select the <u>type</u> of visualization for these variables? (*Hint:* Relate tp the conceptual measure-types for these variables)
    c. How did you improve the graph from its initial ggplot2 output? (*Hint:* Specify some choices of aesthetics, facets or themes that helped improve the visualization of your plot)

    Summary statistics can include-
    a. One-way frequency tables
    b. Two-way frequency tables
    c. Summary tables
    d. Any other summary tables as appropriate

    Visualizations can include
    a. Histograms, bar-graphs, density plots
    b. Box-plots
    c. Scatter-plots, line-plots
    d. Any other plots as appropriate

4. **Preliminary statistical analyses –** Using the takeaways from the previous part, comment which of the relationships are statistically significant. You can do this by performing –
    a. Hypothesis testing – t-tests (one-sample, two-sample), chi-squared tests, correlation tests
    b. Any other techniques as appropriate

    Again, you should relate this to the patterns found in the previous part, particularly –
    a. Why did you select the variables that you chose?
    b. Why did you select the analysis technique for these variables? (*Hint:* Relate to the conceptual measure-types for these variables)
    c. What question are you trying to answer with these statistical analyses?
    d. What conclusion do you draw about the answer to your question? How does this relate to the overall project goal?

5. **Regression analyses -** Using the takeaways from the previous parts, we can now build a regression model, focusing on the factors/determinants of your outcome variable. For this part,
    a. Set up the baseline regression model in R
    b. Clearly state the dependent and independent variables used in your analysis
    c. Clearly state any data aggregations you may have used for running the regression (For eg. – In the Black Friday data, you may choose to look at overall expenditure for users, rather than expenditure at the user-product level. Or in the Flights data, you may choose to look at overall delays for airlines, or airline-airport, rather than the flight-airport level).
    d. Discuss the results of the regression analysis, and clearly state what you can predict using this regression model.
    e. You may extend this analysis by further controlling for omitted variable biases. Mention which variables you are using as controls, and re-do the regression from part (a) after adding these controls. Interpret your findings and note the change in adjusted R-squared/AIC/BIC.
    f. You may also run other regression models on different data aggregation levels. If you choose to do so, state the differences between your different regression models, and the additional benefits of using different data aggregation levels. Are there additional questions you can answer with the other model(s)/aggregations?

6. **Segmentation:**
    a. In your final selected model, are there significant segments based on demographics? Explain.
    b. Can you further the segmentation analysis using a post-hoc segmentation technique? Does this improve your model fit from (5)?

7. **Predictive Analyses –** From the analyses above,
    a. Choose your final prediction problem (For eg – for the Black Friday dataset, predicting the final wallet-spend for an individual, or how much an individual spends in a product category. For the flights data, predicting the average delay at an airport, or an airline, or airport-airline, or airport-airline-month combination. For the Kickstarters data, predicting the likelihood of a project to be successful or the pledged amount). This prediction problem could be related to the regressions you ran earlier in parts (5) and (6).
    b. For this problem, state the data aggregation level you are using. Split the dataset randomly in a 90-10 split for training and validation sets

c. Run at least 15 predictive models on the training set and get the prediction error from the validation set. Show a table with 2 columns, one for the model and the other with the associated prediction error. *Hint: You can choose Root Mean Squared Error (RMSE) or hit rate as the error metric.*

d. Choose the model with best predictive fit, and show its regression summary output.

e. Using your final selected model, predict the mean outcome for the major segments.

   i. *For this part, create a "test" data set. This dataset has observations/rows which correspond to each segment.*

   ii. *This might involve using the cluster membership information from the post-hoc segmentation analysis (using the seg.summ() function to choose the mean/median/mode value for each variable) corresponding to each segment.*

   iii. *For this test data set, predict the mean outcome for each observation/segment.*

f. Using the standard error of prediction from the above prediction analysis, construct the 95% confidence interval for your prediction outcome.

8. **Conclusion –** Summarize your findings for your firm's managers. Discuss what the data patterns and model results indicate, which segments should the managers focus on, and the predicted

While the above points represent basic guidelines, teams have autonomy in choosing which details are most interesting to include in the report. These also involve choosing the key variables for analysis. Brevity is appreciated.

Reports:

There is a LATTE Assignment page, and each group must submit a zipped folder with –

a. The RMarkdown file (.Rmd)
b. The knitted RMarkdown HTML/PDF file (see instructions below)
c. Presentation slides (see instructions below)

Page limit – 25 pages.

Points to note –

1) **File Name:** File-name should indicate member names.

2) **Authorship:** All team-members should have their names under authors in the PDF file. Points will be given only for authors of the document.
3) **Writing, code & output:** Each section will be graded on the quality of writing/explanation, code and the associated output. Points will be deducted if there is insufficient explanation, or if the code/output are not visible in the PDF document.

Specific requirements and guidelines for report:

- You are permitted to supplement your data with other publicly available data if you deem it appropriate.
- If you are using one of the Kaggle-supplied kernels, you must first get the original kernels to run, and cite them in the report. As part of your report and presentation, you should interpret the output of the code.
- You must, in some fashion, make a *substantial modification* to the original kernel (if any) to better respond to the supplier's questions. This might involve selecting different explanatory features, transforming data to accommodate non-linearity, using a different modeling approach, creating a more informative visualization, or other similar changes. Improvements should go beyond simple cosmetic changes or dropping a variable from the model. If you're uncertain whether a particular change is substantial, speak with me.
- Your R Markdown file should contain relevant code as well as narrative explaining the different parts of your investigation, explaining what each code chunk does and why you created the chunk of code. As with earlier assignments, knit the markdown file to a Word doc, then **resave** that document as a pdf for upload to LATTE.
- You **should not** be making any changes to the knitted word document. When I compile the RMarkdown code, I should see the same output as your final PDF document.

Specific requirements and guidelines for presentation slides:

- Slides would be the presentation you make to the company manager. For the sake of your final presentation, imagine I am the company manager with very little data experience. You have to create a deck of slides that explains your analysis simply to the manager.
- Slide breakdown -

Title	1
Introduction/Problem setup	1-2
Slide Outline	1
KEY Descriptive/Summary statistics (Includes tables, plots, handling of missing data)	4-5
Preliminary statistical analysis results	2-3

(Includes hypothesis testing, basic regression analysis)	
Model selection results	2-3
Predictive analysis results	2-3
Summary of results/conclusion	1-2
Appendix/Key assumptions/ Additional results	Optional

- Slides should be pithy and concise. NO long-run sentences!
- Plots and tables should be well-formatted with titles and explanations
- Slides should have the same font and formatting throughout!
- Do NOT post R code on slides! If you want to write out the regression formula, use the Equation .Editor in PowerPoint
- There should NOT be technical details (such as specification of Null/Alternate hypothesis) in the slides. These would be part of your oral presentation, if the question arises during the talk.
- Your submitted presentation would the final version of the slides you would be presenting. Make sure you read them once before submitting.

Under no circumstances should any individual or team in the class copy or share their code with any other individual or team. Code sharing, however minimal, will result in a score of 0 points for all members of any team involved.