MSBA Capstone Exit Exam: Yelp Case Study

ANLY 6900 Assurance of Learning Assessment

This MSBA Exit Exam is designed to evaluate your understanding of the main concepts and learning goals of the MSBA program. In particular, it is designed to help identify areas in which you are especially strong as well as areas in which you may want to consider strengthening your skills. The exit exam is constructed from a number of real data analytics interview tasks, highlighting some of the main ideas of the MSBA program.

For this exit exam, imagine you are a consultant hired by Yelp to provide insights into their business and give suggestions for the online business review industry as Yelp emerges from the COVID-19 pandemic. Yelp has given you a subset of its listings data to analyze, containing approximately 120,000 records and 12 variables pertaining to Yelp online reviews.

Yelp would like you to investigate this sample data set using a variety of analytics methods with the ultimate goal of predicting the average rating (the *stars* variable) for a new Yelp listing. This will help them focus on promoting highly-rated businesses and allocate resources appropriately.

The sample of listings is contained in the data set "MSBA Exit_Yelp.csv". The variables contained in the data set are:

Variable	***Definition***
attributes	A list of attributes selected by a business on the Yelp platform, including any particular options selected by the business (for example, noise level = "average")
business_id	A random business ID assigned to each Yelp business
categories	A list of categories selected by a business on the Yelp platform. Note that business can select up to approximately 40 business categories on the platform
city	The city in which the business is located
hours	The hours the business has entered as open for business
is_open	Whether the business is still currently open and accepting customers (1) or currently closed (0). Note a business that is temporarily closed will be noted as 0 in this data
latitude	The latitude at which the business is located
longitude	The longitude at which the business is located
postal_code	The postal code for the region in which the business is located
review_count	The number of reviews the business has on the Yelp platform
state	The state in which the business is located
stars	The average rating the business has received on the Yelp platform (out of a total of 5 stars)

Using this data, prepare an analysis for Yelp that specifically answers the following questions:

Exploratory Data Analysis

- *Provide a summary of data cleaning issues you have identified in exploring the data. In particular, you should address:*
    - *Are there missing values that could pose a problem? How do you suggest missing data be tackled with these variables/this data set?*
    - *What features of the data require other types of cleaning? (That is, are there issues with the non-missing data that should be cleaned somehow?)*
- *Create a minimum of three (3) data visualizations to illustrate important insights you have been able to derive for the data, including explanations of the visualizations you have created and what they demonstrate.*
- *Consider variables such as "attributes", "categories", and "hours" in this data. Provide a detailed proposal for how Yelp could use the text data tools to create additional variables that could be used for model building. In particular:*
    - *How might regular expressions be helpful here?*
    - *Would more detailed NLP analysis such as LDA be useful? Why or why not?*

Model Development

- *Provide a high-level proposal of at least three (3) models you think could address Yelp's main goal: predicting the average rating of a Yelp business.*
- *Your proposal should include:*
    - *A model chosen specifically for prediction accuracy*
    - *A model chosen specifically for inference/variable insights*
    - *A third model of your choosing*
- *For all models, make sure to give Yelp the model specification as well as a complete description of why you have chosen this model and what you expect to be able to evaluate.*
- *In this section, you should also describe your suggested method for evaluation of the data/models. That is, how will you determine (for each model) the success of the model using this data?*

Machine Learning Analysis

- *Run the models you have proposed in the previous section. For each model, provide Yelp with:*
    - *Any relevant model output*
    - *A complete explanation of the model output*
    - *Any insights or results the model shows*
- *Note that as part of this analysis, you should include the evaluation of the model per your suggestion in the previous section, including any data partitioning or other methods/metrics for evaluation.*
- *This section should also include the text analysis you identified/proposed in your exploratory data analysis.*

Notes:

- **Yelp will ask for both your code and your analysis, so please provide both in an easy-to-follow format! If you are planning to provide one or more code files, make sure the code is commented to clearly connect to your analyses.**
- **Your Exploratory Data Analysis section does not need to be completely comprehensive. However, you should have identified a considerable number of data cleaning issues with this data—it is scraped directly from Yelp, with very limited cleaning!**
- **Note that much of this is generic, high-level discussion of modeling. You do not need to have perfect modeling, perfect data cleaning, etc. However, you should provide clear descriptions and ideas of what you would suggest even if you have difficulty carrying out your suggestions.**
- **For example, if you remove any missing data in order to run your models rather than choosing to do any sort of imputation, you should be very clear either on why you removed the missing data or on what you should have done instead!**