**Exploratory Data Analysis (EDA) and essay**

**Introduction**

This assignment focuses on one of the most important aspects of data science, Exploratory Data Analysis (EDA). Many surveys show that data scientists spend 60-80% of their time on data preparation. EDA allows you to identify data gaps & data imbalances, improve data quality, create better features and gain a deep understanding of your data before doing model training - and that ultimately helps train better models. In machine learning, there is a saying - "better data beats better algorithms" - meaning that it is more productive to spend time improving data quality thanimproving the code to train the model.

This will be an exploratory exercise, so feel free to show errors and warnings that arise during the analysis.

**Dataset**

A Portuguese bank conducted a marketing campaign (phone calls) to predict if a client will subscribe to a term deposit  The records of their efforts are available in the form of a dataset. The objective here is to apply machine learning techniques to analyze the dataset and figure out most effective tactics that will help the bank in next campaign to persuade more customers to subscribe to the bank's term deposit. Download the Bank Marketing Dataset from: <https://archive.ics.uci.edu/dataset/222/bank+marketing>

**Assignment**

1. **Exploratory Data Analysis**  
   Review the structure and content of the data and answer questions such as:
   * Are the features (columns) of your data correlated?
   * What is the overall distribution of each variable?
   * Are there any outliers present?
   * What are the relationships between different variables?
   * How are categorical variables distributed?
   * Do any patterns or trends emerge in the data?
   * What is the central tendency and spread of each variable?
   * Are there any missing values and how significant are they?
2. **Algorithm Selection**Now you have completed the EDA, what Algorithms would suit the business purpose for the dataset. Answer questions such as:
   * Select two or more machine learning algorithms presented so far that could be used to train a model  
     (no need to train models - I am only looking for your recommendations).
   * What are the pros and cons of each algorithm you selected?
   * Which algorithm would you recommend, and why?
   * Are there labels in your data? Did that impact your choice of algorithm?
   * How does your choice of algorithm relates to the dataset?
   * Would your choice of algorithm change if there were fewer than 1,000 data records, and why?
3. **Pre-processing**  
   Now you have done an EDA and selected an Algorithm, what pre-processing (if any) would you require for:  
   * Data Cleaning - improve data quality, address missing data, etc.
   * Dimensionality Reduction - remove correlated/redundant data than will slow down training
   * Feature Engineering - use of business knowledge to create new features
   * Sampling Data - using sampling to resize datasets
   * Data Transformation - regularization, normalization, handling categorical variables
   * Imbalanced Data - reducing the imbalance between classes

**Deliverable**

1. **Essay (minimum 500 words)**  
   Write a short essay summarizing your findings. Explain your selection of algorithms and how they relate to the data and what you are trying to do.  
   *Format: PDF*
2. **Code**  
   This should include your code, as well as the outputs of your code e.g. correlation chart  
   F*ormat: Code should be saved in*[*https://rpubs.com*](https://rpubs.com/)*or*[*https://github.com*](https://github.com/)*. Please provide a link to your code repo in the submission. Please do not submit your code via Google Colab (due to permissioning issues).*

**Q&A:**

1. Python or R  
   Previously this course used only R (as most of your prerequisite courses used R). Due to student demand the course has been extended to cover Python - so you can use the Python textbook and do the assignments in Python.
2. Working individually vs groups  
   Assignments should be submitted individually so credit can be assigned to person doing the work. Any work you copy or leverage should be attributed e.g. from the internet or from another student.
3. RPub and Github accounts  
   Please register for an account if you don't already have one (accounts are free).
4. Google Colab  
   I don't accept Google Colab. 80% of the time Google Colab hasn't been set up to give me permissions to view the code. I cannot grade work I don't have access to.

**Rubric**

|  |  |  |
| --- | --- | --- |
| **Activity** | **Points** | **Requirements** |
| Exploratory Data Analysis (EDA) | 20 | 1. Features (columns) correlation, graph & insights drawn (2) 2. Relationship between variables (this is more of a domain knowledge  (2) question compared to #1 i.e. even if variables are not correlated can they be combined e.g. height & weight combined to be new feature BMI?) (2) 3. Feature distribution & insights drawn (2) 4. Outliers detected; impact and recommendations for handling (2) 5. Categorical variables (2) 6. Patterns/trends in data, and insights drawn (2) 7. Central tendency & spread of variables (2) 8. Missing values & significance (2) 9. Duplicate or inconsistent values identified (2) 10. Does the data align with domain knowledge & expectations (2) |
| Algorithm Selection | 20 | Justification for selection of algorithm (model) based on: 1. Data characteristics, constraints & insight (10) 2. Problem Type (Supervised vs. Unsupervised) (5) 3. Business insight i.e. how does the algorithm align with business goals e.g. interpretability vs accuracy, data availability, speed, scaling, etc. (5)     - I am looking for you to think about business goals, and how you are working towards that |
| Pre-processing | 20 | 1. Data quality & completeness (5)     - addressing missing data     - including handling of outliers & noise 3. Feature distribution & scaling (5)     - features were normalized/standardized     - ensure data distribution aligns with model 4. Feature Selection (5)     - appropriate features selected for predictive power     - handling of categorical variables     - creation of synthetic variables (optional) 4. Domain knowledge & business context (5)     - incorporate business & data insights into above     - ensure features align with real-world, and make sense |
| Code | 20 | 1. Code provided (5) 2. Code quality & completeness (5) 3. Comments & explanation provided (5) 4. Outputs included (5) |
| Essay | 20 | 1. Exploratory data analysis (5)     - Summarize findings & issues     - Conclusions drawn about data 2. Algorithm selection (5)     - selection of algorithm is justified & consistent with EDA 3. Pre-processing (5)     - Issues identified in EDA were addressed     - Changes specific to algorithm were identified & addressed 4. Business insight & recommendations (5)     - Conclusions were reached & justified by actual results/metrics |
|  | **100** |  |