ADTA 5340: Discovery and Learning with Big Data

Thuan L Nguyen, PhD

Final Project

1. Overview

The final project covers all the topics discussed during the course. The materials posted for the class activities should be considered and used for the project.

To work on the final project:

- Downloads a template document from Canvas: **ADTA5340_final_project_template.docx**
- Rename the file as **ADTA5340_final_project_studentname.docx**
- Use the template to submit the student's work on the final project, except for the Python code.

IMPORTANT NOTES:
--) If an MS Word document is specified as the required format of the submitted document, the student should submit it, ***not*** *submit a PDF.*

--) All the submission requirements are expected to be submitted in an MS Word document, except for Python code or being specified otherwise.

--) When discussing a topic or answering a question, the student is expected to provide adequate explanation and supporting details.

IMPORTANT NOTES:
--) For the Python code, the student must write the ***code of each step*** *in* ***one cell*** *of the Jupyter Notebook document, as shown in the lectures. Then, the student must* ***run the code of each cell*** *to* ***show the results of each step*** *in each submitted Jupyter Notebook document.*

--) For Python code in Jupyter Notebook, the student must run the code and submit the Jupyter Notebook document containing the results. The student should ***refrain from copying*** *the results of Python code into the MS Word document.*

2. Data Sets

All the datasets are posted in the Canvas module: …/DATA_SETS

3. PART I: A Strategy to Employ Machine Learning in a Firm (10 Points)

The student is assumed to be the Chief Information/Data Officer (CIO/CDO) of a publicly listed company. To gain a competitive advantage and achieve the business goals, the firm's executive board has decided to employ machine learning (either technology or products) company-wide so that all the departments can take advantage of it and improve their business. The CEO has called a meeting in which he asked the student (CIO/CDO) to implement the board's decision by designing a strategy to employ the technology/products and presenting the plan to the executive board within three weeks.

--) Based on the knowledge and skills that the student has acquired during the course, he/she designs a strategy to employ machine learning technology and products/services in the company.
--) To make it simpler for the student, the company is assumed to have enough financial resources to execute any strategy suggested by the student.
--) It is also assumed the competitors in the same sector are very aggressive in using the new technology/products to enhance their competitive advantage.

--) It is required that the strategy includes – but **not** limited to:

- Considering all the critical aspects of the firms: size, sector, competitors, etc.
- Considering all the major factors that can have a significant impact on the project: technology (which framework, system, etc.), human intellectual capital (staff skills, staff competence, training programs, etc.), etc.
- For the technology, if some technology ecosystem will be deployed, specify which components or sub-systems should be the focus.

SUBMISSION REQUIREMENT PART I:
--) Document the strategy with all the details and supporting information and prepare it to submit to the executive board for consideration.

IMPORTANT NOTES:
--) The student can select any real company as the target for his/her work. Otherwise, the student can imagine a "virtual" company with all the details of a real company – size, number of employees, products & services, etc.

4. PART II: Data Preprocessing (10 Points)

TO-DO
--) Search on the Internet, using Google search or any other approach, to find a dataset in the public domain, i.e., available for use without restrictions.

- This dataset **only** includes text or numeric values and no multimedia content like images or sound files.
- This dataset includes **at least 2000** (two thousand) records.
- This dataset may contain missing values or other issues that may impact the quality of the data.
- This dataset should **not** be any dataset that has been used for class assignments, including the final project.

--) Clean the dataset, i.e., handle missing values, if necessary.

IMPORTANT NOTES:

--) To detect and handle the missing values in the dataset, the student can apply any approach that he/she finds appropriate and comfortable, including writing Python code (if he/she knows how to do it), using tools (if he/she knows any tool), or doing it manually, e.g., open the CSV file with Excel, looking for the missing value and remove the record from the file.

SUBMISSION REQUIREMENT PART II:

--) Write a report on the dataset that includes (but not limited to):

- All the critical information of the dataset, e.g., name, official website, links to download, the data (how many items), the data structure of the data contained in the dataset, list of attributes, the data type of each attribute, which attributes can accept 0 (zero) values and which ones cannot, and so on.
- The quality of the data: Missing values? With which attributes?
- How to handle the missing values?
- Provide a brief summary of a machine learning project that can be done with the dataset.
    - For example: *The dataset can be used to predict … what … based on the following predictors: …*

--) Submit the CSV file of the original dataset.

--) Submit the CSV file of the cleaned dataset (if the cleaning has been done)

IMPORTANT NOTES:

--) Some suggestions for websites to start:

- *Dataset Search (Google): https://toolbox.google.com/datasetsearch*
- *Kaggle: https://www.kaggle.com/datasets*
- *Awesome public datasets:* https://github.com/awesomedata/awesome-public-datasets

IMPORTANT NOTES for the next sections **(PART III, IV, and V):**

--) Before working on any dataset, it is expected that the student has to **perform the exploratory data analysis**.

--) For Exploratory Data Analysis (EDA), **each step** of this analysis must be coded in **one cell**.

--) For Exploratory Data Analysis (EDA), **univariate data visualization**, **each chart** of **each applicable variable** must be displayed in **its own plot**.

IMPORTANT NOTES for the next sections **(PART III, IV, and V):**

--) In some datasets, the first column is real data, not the index column. To force Pandas not to use the first column as the data frame index column, the option “index_col=False” should be included in the Python code to read the dataset. For instance:

```
df = pd.read_csv(filename, names=col_names, index_col= False)
```

5. PART III: Build, Train, Test, and Evaluate ML Models (20 Points)

The dataset **abalone.csv** has the following attributes:

1. Sex
2. Length: mm: Longest shell measurement
3. Diameter: mm : perpendicular to the length
4. Height : mm : with meat in the shell
5. Whole weight : grams : whole abalone
6. Shucked weight : grams : weight of meat
7. Viscera weight : grams : gut weight (after bleeding)
8. Shell weight : grams : after being dried
9. Rings : integer : +1.5 gives the age in years

TO-DO

--) Preprocess the dataset if necessary, including

- Handling missing values
- Handling **abnormal** values, i.e., text values of a numeric attribute such as "3+" that indicates the value of "3 or more."

--) Select a **machine learning model** that has been discussed in the class to work on the dataset.
--) Build, train, and test the model on the dataset with the Python library Scikit-Learn in a Jupyter Notebook document.
--) Make two new records with **reasonable** values of all the predictors of the dataset.
--) Use the trained machine learning model to **predict the age of the abalones** represented by these two new records.
--) Evaluate the model using the **10-fold** cross-validation technique.

--) Report and discuss the results of each significant step:

- Provide an explanation of whether or not the dataset needs preprocessing. If YES, how?
- Provide an explanation in detail of why the model is selected.
- Building the model
- Train the model
- Making up two new records (presenting the value of each attribute)
- Predicting the age of the abalones and interpreting the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT PART III:

IMPORTANT NOTES:
--) It is expected that the student provides all necessary comments for the code in each cell.

--) Show the work in a native Jupyter Notebook document.
--) Each step is coded in one cell.
--) Run the code of each step to show the results.
--) Report and discuss the results of each major step (in the **MS Word** document)
--) Submit the CSV file of the cleaned dataset (if the cleaning has been done)

6. PART IV: Build, Train, Test, and Evaluate ML Models (20 Points)

With nearly 50,000 (fifty thousand) records, the dataset **adult_salary.csv** was collected with the following attributes in a census survey:

1. Age: The age of the individual
2. Emp_type: The type of employer the individual has, e.g. government, military, private, …
3. Fnlwgt: An attribute used only for the census survey purpose- will be **removed**
4. Education: The highest level of education achieved for that individual
5. Education_num: the highest level of education of the individual in the numerical form
6. Marital: The Marital status of the individual
7. Occupation: The occupation of the individual
8. Relationship: The most prominent relationship
9. Race: The race of the individual
10. Sex: The biological sex of the individual
11. Capital_gain: Capital gains recorded
12. Capital_loss: Capital losses recorded
13. Weekly_hours: Number of working hours per week
14. Country: The original country of the individual
15. Income: Whether the person's annual income is **more than** or **less than and equal** to 50,000.00

IMPORTANT NOTES:

*--) The dataset contains **missing values** that are indicated with the **character '?'**.*

TO-DO

--) Preprocess the dataset if necessary, including

- Handling missing values
- Handling **abnormal** values, i.e., text values of a numeric attribute such as "3+" of the attribute "Dependents" in the dataset "loan_approval.csv."

--) Select a **machine learning model** that has been discussed in the class to work on the dataset.

--) Build, train, and test the model on the dataset with the Python library Scikit-Learn in a separate Jupyter Notebook document.

--) Make two new records with **reasonable** values of all the predictors of the dataset.

--) Use the trained machine learning model to **predict the income** (whether the person's annual income is **more than** or **less than and equal** to 50,000.00) represented by these two new records.

--) Evaluate the model using the **10-fold** cross-validation technique.

--) Report and discuss the results of each significant step:

- Provide an explanation of whether or not the dataset needs preprocessing. If YES, how?
- Provide an explanation in detail of why the model is selected.
- Building the model
- Train the model
- Making up two new records (presenting the value of each attribute)
- Predicting the income and interpreting the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT PART IV:

IMPORTANT NOTES:

--) It is expected that the student provides all necessary comments for the code in each cell.

--) Show the work in a native Jupyter Notebook document.
--) Each step is coded in one cell.
--) Run the code of each step to show the results.
--) Report and discuss the results of each major step (in the **MS Word** document)
--) Submit the CSV file of the cleaned dataset (if the cleaning has been done)

7. PART V: Build, Train, Test, and Evaluate ML Models (20 Points)

The dataset **car_evaluation.csv** was collected with the following attributes:

1. Price: Buying price
2. Maintenance: Maintenance cost
3. Doors: Number of doors
4. Passengers: Number of passengers
5. Luggage: Size of luggage boot
6. Safety: Estimated safety of the car
7. Evaluation: Evaluation of the car

The dataset can be used to predict car evaluation that can be classified as unacceptable, acceptable, good, or very good.

TO-DO

--) Preprocess the dataset if necessary, including

- Handling missing values
- Handling **abnormal** values, i.e., text values of a numeric attribute such as "3+" of the attribute "Dependents" in the dataset "loan_approval.csv."

-) Select an **unsupervised machine learning model** discussed in the class to work on the dataset.
--) Build, train, and test the model on the dataset with the Python library Scikit-Learn in a separate Jupyter Notebook document.

--) Make two new records with **reasonable** values of all the predictors of the dataset.

--) Use the trained machine learning model to **predict** the **cluster** to which the data points represented by these two new records belong.
--) Report and discuss the results of each significant step:

- Provide an explanation of whether or not the dataset needs preprocessing. If YES, how?
- Provide an explanation in detail of why the model is selected.
- Build the model
- Train the model
- Make two new records (presenting the value of each attribute)
- Predict the cluster to which the data points belong and interpret the results.

SUBMISSION REQUIREMENT PART V:

IMPORTANT NOTES:

--) It is expected that the student provides all necessary comments for the code in each cell.

--) Show the work in a native Jupyter Notebook document.
--) Each step is coded in one cell.
--) Run the code of each step to show the results.
--) Report and discuss the results of each major step (in the **MS Word** document)
--) Submit the CSV file of the cleaned dataset (if the cleaning has been done)

8. PART VI: Evaluate and Compare Machine Learning Models (20 Points)

8.1 Regression Models: Linear Regression vs. Decision Tree (CART) Regression

8.1.1 R-Square

TO-DO

--) Train both models on the same dataset:

- Any dataset that has at least 2000 records.
- The dataset can be one of those provided in the final project or one found by the student.
  - The dataset **must** be preprocessed, if necessary, before being used.

--) Make observations and compare the values of **R^2** obtained in the results.

SUBMISSION REQUIREMENT PART VI #1:

--) Write a detailed report on the quality of these models based on the observations.
--) The report should include all the details about the dataset and how to preprocess it.
--) Submit the original dataset and the cleaned dataset (if preprocessing is done, and the datasets have not been submitted for any of the previous submission requirements.)

8.1.2 Prediction

TO-DO

--) Make a new set of predictors for the dataset.
--) Use both models to predict the new record

SUBMISSION REQUIREMENT PART VI #2:

- Write a report on the predictions made by these two models on the same new data records, focusing on whether they are the same or not.
- If the results of predictions are not the same, use the above report (**comparing R^2 values**) to make a preliminary guess about which model may predict more accurately.

8.1.3 K-Fold Cross-Validation

TO-DO

--) Evaluate each model using 10-fold cross-validation.

SUBMISSION REQUIREMENT PART VI #3:

- Write a report on the average error estimations of these two models, focusing on whether they are the same or not.
- Use these values to evaluate the quality of the models. If the results are not the same, what is the difference?
- Make a conclusion, if possible, on which model has higher quality in predicting outcomes and should be selected as the model to predict the new data.
- Based on the above conclusion, write a report on the predictions that should be made on the new data records

8.2 Classification Models: Logistic Regression vs. K-Nearest Neighbors

8.2.1 Accuracy Level

TO-DO

--) Train both models on the same dataset:

- Any dataset that has at least 2000 records.
- The dataset can be one of those provided in the final project or one found by the student.
  - The dataset **<u>must</u>** be preprocessed, if necessary, before being used.

--) Make observations and compare the accuracy levels of both models obtained in the results.

SUBMISSION REQUIREMENT PART VI #4:

--) Write a detailed report on the quality of these models based on the accuracy level of each model.
--) The report should include all the details about the dataset and how to preprocess it.
--) Submit the original dataset and the cleaned dataset (if preprocessing is done, and the datasets have not been submitted for any of the previous submission requirements.)

8.2.2 Prediction

TO-DO

--) Make a new set of predictors for the dataset.
--) Use both models to predict the new record

SUBMISSION REQUIREMENT PART VI #5:

- Write a report on the predictions made by these two models on the same new data records, focusing on whether they are the same or not.
- If the results of the predictions are not the same, use the above report (comparing the accuracy levels) to make a preliminary guess about which model may predict more accurately.

8.2.3 K-Fold Cross-Validation

TO-DO

--) Evaluate each model using 10-fold cross-validation.

SUBMISSION REQUIREMENT PART VI #6:

- Write a report on the average error estimations of these two models, focusing on whether they are the same or not.
- Use these values to evaluate the quality of the models. If the results are not the same, what is the difference?
- Make a conclusion, if possible, on which model has higher quality in predicting outcomes and should be selected as the model to predict the new data.
- Based on the above conclusion, write a report on the predictions that should be made on the new data records.

9. Grading Criteria

The final project, including the final presentation, is graded based on the following grade components:

1. **Final project report (PART I – PART VI) (100 Points): 70%**
2. **Final project: Online – On Canvas (PART VII) (100 Points): 30%**

9.1 Final Project Report (100 Points)

The student must **submit the final project report** (an MS Word document) along with the following documents **(total 6 documents,** including the report**):**

1. A CSV file of the original dataset for PART II
2. A CSV file of the cleaned dataset for PART II
3. A Jupyter Notebook document for PART III (in its native format)
4. A Jupyter Notebook document for PART IV (in its native format)
5. A Jupyter Notebook document for PART V (in its native

format) The final project report includes the following sections:

- PART I: A Strategy to Employ Machine Learning in a Firm (10 Points)
- PART II: Data Preprocessing (10 Points)
- PART III: Build, Train, Test, and Evaluate ML Models (20 Points)
- PART IV: Build, Train, Test, and Evaluate ML Models (20 Points)
- PART V: Build, Train, Test, and Evaluate ML Models (20 Points)
- PART VI: Evaluate and Compare Machine Learning Models (20 Points)

9.2 Final Project: Online – On Canvas (100 Points)

The student takes **Final Project: PART VII online on Canvas** on. **Final Project: PART VII**:

- Consisting of **short questions and answers**.
- Being **comprehensive**: It covers all the topics discussed during the semester.
- **OPEN BOOK – OPEN NOTES**: The student can use any class lectures and notes.

IMPORTANT NOTES:
*--) The **correct answers** to each question must be the **contents of the lecture documents** posted on Canvas or **materials discussed in weekly online class meetings**, i.e., **not** the materials that can be found on the Internet or anywhere else.*

IMPORTANT NOTES:
*--) The student must **complete PART VII in 60 minutes (one hour)** although he/she **can start** working on the assignment **any time between 8:00 AM and 10:00 PM on***

10. HOWTO Submit

10.1 Final Project Report and All Related Documents (PART I – PART VI)

The student must submit all required documents – Final Project Report and Jupyter Notebook documents – as attachments to a UNT email that is sent to the instructor .

The subject of the email must be: "ADTA 5340: Final Project – Submission."

IMPORTANT NOTES:
*--) Due to the limited time for grading and posting the grades as required by the Registrar's Office, **no late submission** is accepted.*