ADTA 5340: Discovery and Learning with Big Data

Thuan L Nguyen, PhD

Final Project

1. Overview

The final project covers all the topics that have been discussed during the course. The materials in any format that have been posted for the class activities should be considered and used for the project. Additionally, the student can use any other source of information that he/she can gather.

The student is required to create an MS Words document named "**ADTA5340_final_project.docx**" that will contain all his/her work, except for the Python coding.

IMPORTANT NOTES:
--) The student can use any source of information that he/she considers the best fit for his/her work on the project.
--) The sources can be from class lectures, assignments, etc., or from any other sources
--) Images can include the screenshots that the student has taken while working on the class assignments.
--) Screenshots without details of explaining what they are and what they are for are considered incomplete.

IMPORTANT NOTES:
--) All students are free to discuss with their classmates while working on the final project.
--) However, the final project is an individual assignment. All the submitted documents for the final project are the work done only by the student.

IMPORTANT NOTES:
*--) If an MS Words 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 Words document, except for Python code, or being specified otherwise.

*--) For Python code in Jupyter Notebook, the student is required to run the code and submit the **native** (not PDF) Jupyter Notebook document that contains the results. The student should **not** copy the results of Python code into the MS Words 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 (15 Points)

It is assumed that the student is the Chief Information/Data Officer (CIO/CDO) of a publicly listed company. To gain competitive advantage and achieve the business goals, the executive board of the firm has decided to employ the 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 present the strategy 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 the machine learning technology and products/services in the company.
--) To make it simpler for the student, it is assumed that the company has enough financial resource 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 #1:
--) Document the strategy with all the details and supporting information and make it ready 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: Big Data, Artificial Intelligence, and Machine Learning (15 Points)

SUBMISSION REQUIREMENT #2:

- **Question 2.1**:

  Do research and discuss (**Min: 3 pages, including images**) the history of artificial intelligence until now, focusing on the recent advancement of the field.

- **Question 2.2**:

  Select three different sectors of the U.S. economy, do research, and discuss (**Min: 3 pages including images**) the impacts of **big data** and **machine learning** on **each** of them.

IMPORTANT NOTES:
--) The student can select any sector in which he/she is interested. For example, he/she can choose high-tech, retail, and transportation, or healthcare, education, and manufacturing, to name a few.

- **Question 2.3**:

  Discuss **in detail** the three major styles of learning in machine learning: (1) Supervised Learning, (2) Unsupervised Learning, and (3) Semi-Supervised Learning.

5. PART III: 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** contains text or numeric values, no multimedia contents like images or sound files.
- This dataset contains **at least 2000** (two thousand) records.
- This dataset may contain missing values

--) 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 #3:

--) 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 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 IV, V, VI,** and **VII):**

--) 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 IV, V, VI,** and **VII):**

--) 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)
```

6. PART IV: Machine Learning: Supervised (15 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+” of the attribute “Dependents” in the dataset “loan_approval.csv.”

--) Select a supervised machine learning model to work on the dataset.
--) Build, train, and test the model on the dataset using the supervised learning style with Python library Scikit-Learn in a Jupyter Notebook document.
--) Makeup 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.

--) **Add a section** to the above MS Words document ("**ADTA5340_final_project.docx**") to 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 interpret the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT #4:

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

7. PART V: Machine Learning: Supervised (15 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 loses 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 supervised machine learning model to work on the dataset.

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

--) Makeup 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.

--) **Add a section** to the above MS Words document ("**ADTA5340_final_project.docx**") to 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 interpret the results.
- Evaluating the model
- Interpret the prediction results again based on the results of evaluating the model

SUBMISSION REQUIREMENT #5:

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 Words** document)

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

8. PART VI: Machine Learning: Unsupervised: (15 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."

--) Build, train, and test the model on the dataset using K-Means, an unsupervised learning technique, with Python library Scikit-Learn in a separate Jupyter Notebook document.

--) Makeup 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.

--) **Add a section** to the above MS Words document ("**ADTA5340_final_project.docx**") to 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
- Makeup two new records (presenting the value of each attribute)
- Predict the cluster to which the data points belong and interpret the results.

SUBMISSION REQUIREMENT #6:

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

9. PART VII: Evaluate and Compare Machine Learning Models (15 Points)

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

9.1.1 R-Square

TO-DO

--) Train both the 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 # 7.1

--) 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.)

9.1.2 Prediction

TO-DO

--) Makeup a new set of predictors for the dataset.
--) Use both the models to make a prediction on the new record

SUBMISSION REQUIREMENT # 7.2

- 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, using the above report (**comparing R^2 values**) to make a preliminary guess about which model may predict more accurately.

9.1.3 K-Fold Cross-Validation

TO-DO

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

SUBMISSION REQUIREMENT # 7.3

- 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 make a predictions on the new data.
- Based on the above conclusion, write a report on the predictions that should be made on the new data records

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

9.2.1 Accuracy Level

TO-DO

--) Train both the 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 accuracy levels of both the models obtained in the results.

SUBMISSION REQUIREMENT # 7.4

--) 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.)

9.2.2 Prediction

TO-DO

--) Makeup a new set of predictors for the dataset.
--) Use both the models to make a prediction on the new record

SUBMISSION REQUIREMENT # 7.5

- 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, using the above report (comparing the accuracy levels) to make a preliminary guess about which model may predict more accurately.

9.2.3 K-Fold Cross-Validation

TO-DO

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

SUBMISSION REQUIREMENT # 7.6

- 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 make a predictions on the new data.
- Based on the above conclusion, write a report on the predictions that should be made on the new data records.

10. PART VIII: Final Presentation Videos: YouTube Links

The student is required to submit the video(s) to **YouTube** as discussed in the document: "howto_submit_videos_to_youtube.pdf" posted in the module …/FINAL_PROJECT.

SUBMISSION REQUIREMENT #8:

--) Provide all the YouTube links of the submitted final presentation videos

11. Grading Criteria

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

1. Final project report:	**70%**
2. Final project presentation:	**20%**
3. Final project peer-evaluation:	**10%**

11.1 Final Project Report (100 Points)

The student is required to **submit the final project report** (an MS Words document) along with the following documents **(totally 6 documents** including the report**):**

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

The final project report as an MS Words contains the following sections:

11.1.1 PART I: A Strategy to Employ Machine Learning in a Firm (15 Points)

Submit Submission Requirement #1

11.1.2 PART II: Big Data, Artificial Intelligence, and Machine Learning (15 Points)

Submit Submission Requirement #2

11.1.3 PART III: Data Preprocessing (10 points)

Submit Submission Requirement #3

11.1.4 PART IV: Machine Learning: Supervised (15 points)

Submit Submission Requirement #4

11.1.5 PART V: Machine Learning: Supervised (15 Points)

Submit Submission Requirement #5

11.1.6 PART VI: Machine Learning: Unsupervised (15 Points)

Submit Submission Requirement #6

11.1.7 PART VII: Evaluate and Compare Machine Learning Models (15 Points)

Submit Submission Requirement #7

11.1.8 PART VIII: Final Presentation Videos: YouTube Links

Submit Submission Requirement #8

11.2 Final Project Presentation (100 Points)

11.2.1 Overview

The student is required to make **a video of 30 minutes** to present his/her final project report to the whole class. A video shorter than **27** minutes is considered too short; one longer than **33** minutes is considered too long. Points will be deducted from either case.

The contents of the video should cover all the sections in the final project.

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

1. Contents of the presentation: 70%
   a. i.e., How many % of the final project report is covered in the video?

2. Style of the presentation: 30%
   a. i.e., Does the student present the topics well, clearly and interestingly?

IMPORTANT NOTES:

*--) The student can create **2** or **3 videos** of which **the total length is about 30 minutes** instead of one video of 30 minutes if he/she chooses to do so.*

*--) The student **cannot** submit more than 3 videos for his/her final project presentation.*

IMPORTANT NOTES:
--) It is strongly recommended that all the students should create a separated document, e.g., PowerPoint slides, and use it to make the final project presentation in the video.

--) If the student creates a separated document, e.g., PowerPoint slides, and use it to make the final presentation in the video, he/she is required to submit this document along with the final project report.
IMPORTANT NOTES:
--) The student can use any software tool or online service to create the video.
--) If he/she wants to look for an online service but not sure which one provides a decent service, the student can try: screencast-o-matic.com

11.2.2 Video format

The submitted video should be an **MP4** file that is the format supported (and even the default one) by almost all the software tools and online services to create videos.

11.2.3 Naming final presentation videos

If the student submits only one video to make the final presentation, the video should be named as follows: final_project_presentation_video_<full name>.mp4

For example: final_project_presentation_video_JohnSmith.mp4

If the student **submits more than one videos**, he/she is required to **name the videos in their order**.

For instance:
final_project_presentation_video_JohnSmith_1.mp4
final_project_presentation_video_JohnSmith_2.mp4
final_project_presentation_video_JohnSmith_3.mp4

11.2.4 Submitting Final Presentation Videos

The student is required to submit the video(s) to **YouTube** as discussed in the document: "howto_submit_videos_to_youtube.pdf" posted in the module …/FINAL_PROJECT.

11.3 Final Project Peer Evaluation (100 Points)

The student's final project (the final project presentation, the project report, and all the submitted documents) is graded by the instructor. Each student will receive the instructor's feedback.

Additionally, as a way to help each student find out how the classmates think about his/her work on the final project after watching the video, the student also receives **peer-evaluation opinions** from three **anonymous** classmates.

To submit the final project, the instructor will provide each student with a OneDrive folder to upload the final project video to present his/her work on the project. By the due date to submit the video, only the student can access this folder. When the due date of submission has passed, all the submitted videos will be posted for the whole class to access and watch.

Each student is recommended to watch all the classmates' videos. He/she is also required to **provide peer-evaluation opinions** on the videos of **three classmates** that are randomly assigned by the instructor. The evaluation will be done based on the instructor's guideline. It is strongly advised that the **names of the classmates** who will receive the peer-evaluation opinions are **strictly confidential**. The student provides the peer-evaluation opinions by **filling out a form** and sending it back to the instructor by the due date.

12. HOWTO Submit

12.1 Final Project Report and All Related Documents

The student is required to submit all the required documents – Microsoft Words and Jupyter Notebook documents – as attachments to a UNT email that is sent to the instructor (Thuan.Nguyen@unt.edu)

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

Due date & time: 8:00 AM – Monday 12/02/2019

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

12.2 Final Project Presentation: Videos and All Related Documents

The student is required to submit the video(s) to **YouTube** as discussed in the document: "howto_submit_videos_to_youtube.pdf" posted in the module …/FINAL_PROJECT.

Due date & time: 8:00 AM Monday 12/02/2019

IMPORTANT NOTES:
*--) Due to the limited time for classmates to access and watch the presentation videos as required, **no late submission** is accepted.*

12.3 Final Project Peer-Evaluation Forms

The student is required to submit his/her peer-evaluation forms on three classmates that are sent to the instructor (Thuan.Nguyen@unt.edu) as attachments to a UNT email.

The subject of the email must be: "ADTA 5340: Peer Evaluation Forms– Submission."

Due date & time: 8:00 AM Wednesday 12/04/2019

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

IMPORTANT NOTES:
*--) It is expected that the student submits **three peer-evaluation forms** in the MS Words format, one for each evaluated classmate.*