Final Project
AI Algorithms
20 Points
Group project: Maximum group members-2
Due Dates:

In-class presentation: April 17, (5 points)
Submission to DC Connect: April 17, 11.59 PM (15 points)

Mr. John Hughes has been reviewing the **house_price.csv** dataset. The dataset shows the housing price in a city say "XYZ" over a period of time.

The dataset contains the following variables:

Independent Variables

id - Unique ID for each home sold
date - Date of the home sale
bedrooms - Number of bedrooms
bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living - Square footage of the apartments interior living space
sqft_lot - Square footage of the land space
floors - Number of floors
waterfront - A variable for whether the apartment was overlooking the waterfront or not
view - An index from 0 to 4 of how good the view of the property was
condition - An index from 1 to 5 on the condition of the apartment,
grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
sqft_above - The square footage of the interior housing space that is above ground level
sqft_basement - The square footage of the interior housing space that is below ground level
yr_built - The year the house was initially built
yr_renovated - The year of the house's last renovation
zipcode - What zipcode area the house is in
sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors

sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

Dependent Variables

price - Price of each home sold

The Ask:

Build at least 5 ML models to predict the housing price.

Below is the project template that you can use in your machine learning projects in Python.

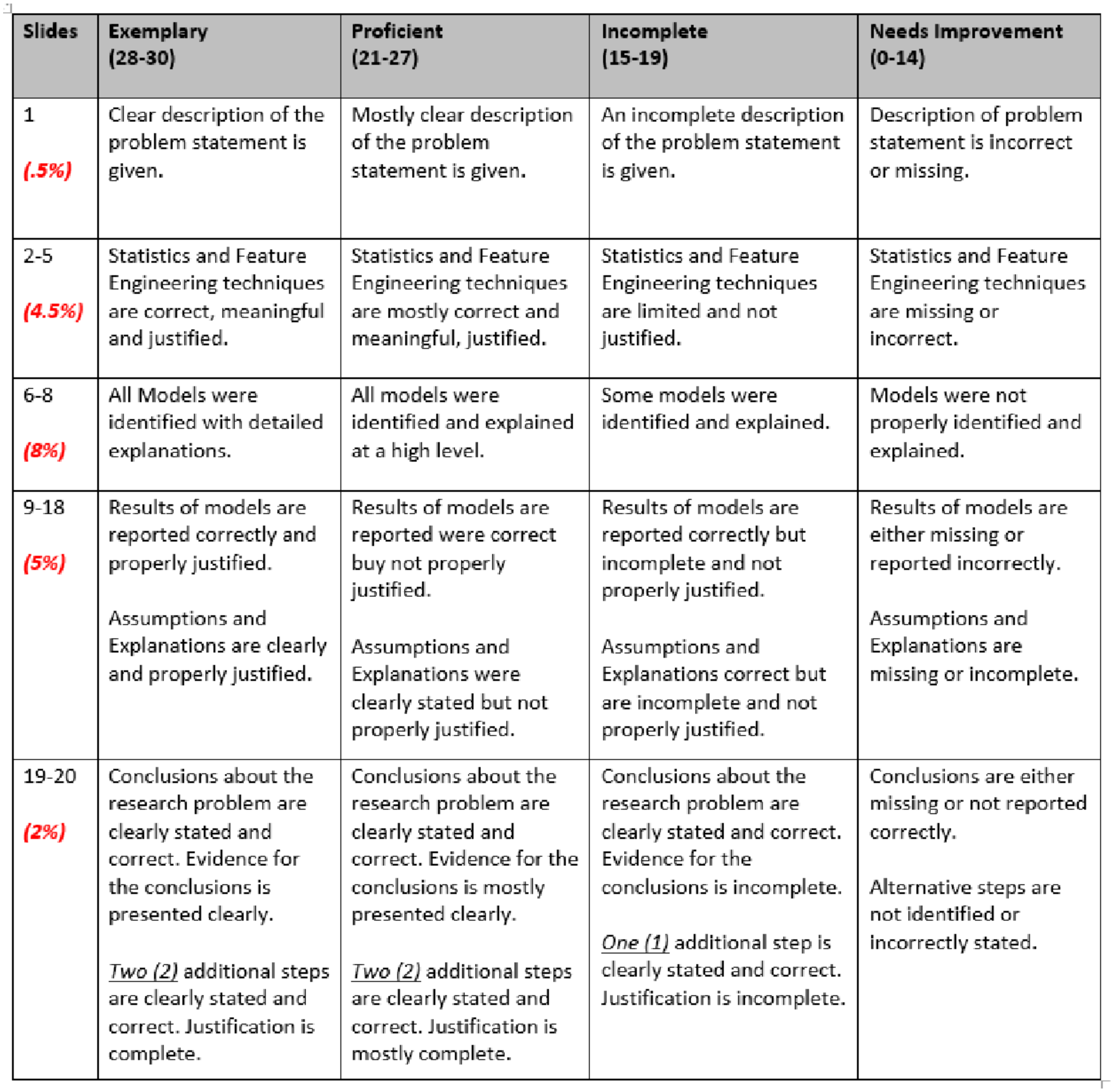
1. Prepare Problem
    a) Load libraries
    b) Load dataset
2. Summarize Data/ Exploratory Data Analysis
    a) Descriptive statistics
    b) Data visualizations
3. Prepare Data
    a) Data Cleaning
    b) Feature Selection
    c) Data Transforms
4. Model Building and Evaluate Algorithms
    a) Split-out validation dataset
    b) Test options and evaluation metric
    c) Spot Check Algorithms
    d) Compare Algorithms
5. Improve Accuracy
    a) Algorithm Tuning
    b) Ensembles
6. Finalize Model
    a) Predictions on validation dataset
    b) Create standalone model on entire training dataset
    c) Save model for later use

7. Present your analysis of each of the above steps. Make conclusions and recommendations.

Use **seed = 7 and random_state= seed when appropriate.**

Note: Follow the best practices discussed throughout our classes.

Final Term Project Rubric



Slides	Exemplary (28-30)	Proficient (21-27)	Incomplete (15-19)	Needs Improvement (0-14)
1 *(.5%)*	Clear description of the problem statement is given.	Mostly clear description of the problem statement is given.	An incomplete description of the problem statement is given.	Description of problem statement is incorrect or missing.
2-5 *(4.5%)*	Statistics and Feature Engineering techniques are correct, meaningful and justified.	Statistics and Feature Engineering techniques are mostly correct and meaningful, justified.	Statistics and Feature Engineering techniques are limited and not justified.	Statistics and Feature Engineering techniques are missing or incorrect.
6-8 *(8%)*	All Models were identified with detailed explanations.	All models were identified and explained at a high level.	Some models were identified and explained.	Models were not properly identified and explained.
9-18 *(5%)*	Results of models are reported correctly and properly justified. Assumptions and Explanations are clearly and properly justified.	Results of models are reported were correct buy not properly justified. Assumptions and Explanations were clearly stated but not properly justified.	Results of models are reported correctly but incomplete and not properly justified. Assumptions and Explanations correct but are incomplete and not properly justified.	Results of models are either missing or reported incorrectly. Assumptions and Explanations are missing or incomplete.
19-20 *(2%)*	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is presented clearly. *Two (2)* additional steps are clearly stated and correct. Justification is complete.	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is mostly presented clearly. *Two (2)* additional steps are clearly stated and correct. Justification is mostly complete.	Conclusions about the research problem are clearly stated and correct. Evidence for the conclusions is incomplete. *One (1)* additional step is clearly stated and correct. Justification is incomplete.	Conclusions are either missing or not reported correctly. Alternative steps are not identified or incorrectly stated.

Submission Format

In the DC Connect, post **the ran jupyter notebook file**. Any submission other than the format of a notebook file will be graded to zero.

Academic Integrity and Late submission:

Assignments are due by the due date announced in class and posted on DC Connect. At his or her own discretion, and depending on the nature of the assignment, each professor will provide a facility for the submission of late assignments up to a maximum of 72 hours after the assignment due date. All allowed late submissions will be assessed a penalty of 25% of the total possible grade for the assignment. Assignments should be submitted on time, on a regular basis, to enable you to stay on track within the class.

Any violation of academic integrity will not be accepted and will be given a grade of zero (0) and reported. Find more information on academic integrity here https://durhamcollege.ca/mydc/learning-resources/academic-integrity